Université de Montréal

Méthodes à noyaux appliquées à la gestion de portefeuille

par
Christian Dorion
Département d'informatique et de recherche opérationnelle
Faculté des arts et sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maîtrise ès Science
Informatique

Mai 2004

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé:

Méthodes à noyaux appliquées à la gestion de portefeuille

présenté par:

Christian Dorion

a été évalué par un jury composé des personnes suivantes:

Felisa J. Vázquez-Abad
(président-rapporteur)

Yoshua Bengio
(directeur de recherche)

Patrice Marcotte
(membre du jury)

Mémoire accepté le:
31 août 2004

Résumé

L'apprentissage automatique est connue pour être un outil fiable et puissant pour la gestion de portefeuilles boursiers (Weigend, Abu-Mostafa et Refenes 1997). Très souvent, on utilise des réseaux de neurones pour effectuer une régression sous le coût quadratique afin d'obtenir des prédictions sur les séries de rendements. La gestion de portefeuille est ensuite déléguée à un modèle économique classique.

Depuis quelques années, certains résultats autant théoriques que pratiques tendent à démontrer que, plutôt que de reléguer le modèle statistique au rôle d'intermédiaire, on gagne à optimiser un modèle sous un critère financier afin d'apprendre directement les positions à prendre sur le marché (Bengio 1997; Chapados 2000).

Avec le développement des méthodes à noyaux, certains auteurs ont cherché à remplacer les réseaux de neurones par des modèles à noyaux dans la prédiction de séries financières (Kivinen, Smola et Williamson 2002). Toutefois, on ne confine encore les modèles qu'à un rôle de prédiction, laissant la décision pour la théorie économique.

Nous suggérons donc un modèle à noyaux qui assume la prise de décision. Nous ne supposons aucun oracle fournissant les suites optimales de positions

à prendre, sortant ainsi légèrement du cadre de régression classique pour se rapprocher des méthodes de maximisation de la vraisemblance.

Mots-clés : Apprentissage machine, Méthodes à noyaux, Critère financier, Utilité financière, Gestion de portefeuilles boursiers, Validation séquentielle

Abstract

Machine learning proved to be a reliable and powerful tool in financial portfolio management (WEIGEND, ABU-MOSTAFA et REFENES 1997). In many cases, one trains, under a mean square error criterion, a neural network to predict financial time series, such as returns series. The portfolio decision is then forwarded to a classic portfolio management model.

In the past few years, theoretical and practical results have tended to show that one would gain from training under a financial training criterion rather than a prediction criterion, getting the learner to decide the investor's positions directly (BENGIO 1997; CHAPADOS 2000).

With the rise of kernel methods, some authors attempted to replace neural networks by kernel-based models for predictions of financial time series (KIVINEN, SMOLA et WILLIAMSON 2002). However, the model is still kept in an intermediate role, the decision still being forwarded to economic models.

We suggest a kernel-based portfolio management model without supposing that any oracle provides us with the optimal portfolio trajectories. Therefore, we go beyond the classical regression setting to deviate slightly towards a decision-theoretic framework.

Keywords : Machine Learning, Kernel Methods, Financial Criterion, Financial Utility, Quantitative Trading, Sequential Validation

Table des matières

Liste des figures

Liste des tableaux

À mes parents, mes grands-parents,

mon petit frère et mes petites sœurs

Remerciements

Les deux dernières années furent pour moi une occasion non seulement d'apprendre, mais aussi de grandir. Évidemment, si je pouvais m'y reprendre, j'attaquerais probablement l'expérience de la maîtrise avec une discipline et, surtout, dans un état d'esprit... différents. Or, comme on ne vit qu'une seule fois, je me retrouve ici, deux ans plus tard, deux ans plus grand, et je me dois de remercier ceux qui m'ont supporté dans cet apprentissage.

D'abord, les gens du LISA, communauté d'un esprit scientifique d'une indéniable grandeur. Merci, entres autres, à Yoshua pour ses judicieux conseils et son infatigable présence, surtout dans les derniers milles de la rédaction de ce mémoire. À Réjean pour son inaltérable bonne humeur malgré mes mille et une questions et à Nicolas, un mentor sans qui le monde de la finance ne serait encore pour moi qu'une énigme.

Aussi, dans cette sphère du savoir qu'est le DIRO, certaines personnes ont su m'éclairer d'une lumière toute particulière. Mes remerciements les plus sincères à Patrice, sans doute le meilleur pédagogue qu'il m'ait été donné de croiser dans mon parcours universitaire, pour m'avoir transmis *une vision* de l'optimisation et m'avoir permit de goûter à l'enseignement. Merci aussi à Michel pour la confiance qu'il a investie en moi en me permettant le seconder et pour les discussions sur notre réalité prévisible. Enfin, un énorme merci à Felisa, l'enseignante mais surtout l'amie, qui, depuis le début, m'a porté un intérêt bienveillant : tu as été pour moi une guide indispensable.

Évidemment, il me reste aussi des remerciements pour ma famille et mes amies. Merci à ma mère, mon père et mes grands-parents ; si le doute tentait de s'installer, je n'avais qu'à penser à vous pour me donner un second souffle. Mais parfois, malheureusement, même le second ne suffit plus. C'est dans de tels moments que l'on est à même d'apprécier le soutient indéfectible d'une amie sincère. Merci Véro, ma réparatrice.

Un merci infini à ma petite sœur, celle qui a souffert mon caractère exécrable dans ses pires moments et qui s'est joint à moi plus d'une fois pour grogner sur les affres de la maîtrise. Ta rédaction s'en vient et moi, je me prépare au pire... Merci Sapha d'avoir toujours été là pour moi ; je ne l'oublierai jamais.

Enfin, un merci tout spécial à la femme derrière le petit homme ; sans toi Kim je n'y serais peut-être jamais arrivé. Crois-moi, tu en fais de la magie.

Je me dois de souligner l'appui financier du CRSNG, organisme subventionnaire grâce auquel la poursuite d'études supérieures est beaucoup plus accessible. Merci.

Partie I

INTRODUCTION

Depuis plus d'un demi-siècle, le choix des biens et des proportions dans lesquelles un individu doit investir, pour faire fructifier ses avoirs tout en respectant son profil de risque, est au centre des recherches de bien des économistes. Ce problème s'incrit dans le cadre de la *théorie moderne du portefeuille.*

Ce mémoire présente un modèle de gestion d'un portefeuille boursier dont l'inspiration, tirée de modèles économiques classiques, se retrouve plongée dans le monde de l'apprentissage automatique moderne.

Nous considérons un modèle à temps discret, c'est à dire dans lequel une *période* (un jour ou un mois, par exemple) s'écoule entre deux prises de décision. À chaque période, le modèle suggérera la composition (un vecteur) du portefeuille idéal pour la période donnée et il sera du domaine de l'évaluation de performance de considérer ce portefeuille dans la réalité du marché. Notre objectif sera de maximiser le compromis moyenne-variance du rendement mensuel moyen de notre stratégie de gestion de portefeuille.

Afin d'éclaircir les enjeux sous-jacents à ce problème, la **Partie II** introduit d'abord les enjeux et modèles économiques considérés dans ce mémoire. Suivra un apperçu du fascinant monde de l'apprentissage automatique (**Partie III**). Finalement, une fois présentés les domaines concernés par notre modèle, nous en élaborerons le détail (**Partie IV**). Mais, avant tout, suit un bref résumé des notations utilisées dans ce mémoire.

CHAPITRE 1

Notations et rappels

Dans ce mémoire, nous aurons recours à certaines notations qu'il vaut la peine d'établir. Aussi cette section se veut-elle un bref survol de ces diverses notations, entremêlées de quelques rappels qui seront utiles pour la suite de la lecture.

1.1 Espaces vectoriels

Nous dénoterons la **droite réelle** par $\mathbb{R}$ et l'ensemble des points à n dimensions réelles par $\mathbb{R}^n$.

Définition 1.1 *Un* **espace vectoriel** $\mathcal{V}$ *sur* $\mathbb{K}$ *est un ensemble fermé sous l'addition et la multiplication, tel que pour tous* **vecteurs** $x, y, z \in \mathcal{V}$ *et tous scalaires* $a, b \in \mathbb{K}$ *on observe les propriétés suivantes :*

(i) Commutativité :
$x + y = y + x$

(ii) Associativité de l'addition vectorielle :
$(x + y) + z = x + (y + z)$

(iii) Existence d'un nul additif :
$\exists \mathbf{0} \in \mathcal{V}$ *tel que* $x + \mathbf{0} = \mathbf{0} + x = x$

(iv) Existence d'un inverse additif :
$\exists (-x) \in \mathcal{V}$ *tel que* $x + (-x) = \mathbf{0}$

(v) Associativité de la multiplication scalaire :
$a(bx) = (ab)x$

(vi) Distributivité de la somme de scalaires :
$(a + b)x = ax + bx$

(vii) Distributivité de la somme vectorielle :
$a(x + y) = ax + ay$

(viii) Existence d'un scalaire neutre multiplicatif :
$\exists 1 \in \mathbb{K}$ *tel que* $1x = x$

Nous mentionnerons ici le cas particulier de $\mathbb{R}^n$, l'**espace vectoriel euclidien** de dimension n. Pour tout point x de $\mathbb{R}^n$ nous considérerons $[x]_i$ comme étant le i^e élément du vecteur x, tandis que x_1 et x_2 dénoterons 2 vecteurs différents.

Nous considérerons un **vecteur** x comme étant un vecteur colonne, et nous utiliserons la notation x^{T} pour marquer la **transposition** de x en vecteur ligne. Finalement, lorsque nous traiterons de vecteurs, les notations $x > \lambda$, où λ est une constante, et $x > y$ devront être interprétées composante par composante, c'est à dire qu'elles sont respectivement équivalentes à dire que $[x]_i > \lambda, \forall i$ et $[x]_i > [y]_i, \forall i$. Il va de soit que le même raisonnement s'applique pour les autres relations d'ordre.

Définition 1.2 *Soit* $\mathrm{S} = \{x_1, \ldots, x_m\}$ *des éléments d'un espace vectoriel* $\mathcal{V}$. *Le sous-espace vectoriel* $\mathrm{span}\{\mathrm{S}\} \subset \mathcal{V}$ *engendré par* S *est*

$$\mathrm{span}\{\mathrm{S}\} = \{x \in \mathcal{V} \,|\, \exists \gamma_i \in \mathbb{R}, i = 0, \ldots, m, \text{ tels que } x = \sum_{i=1}^{m} \gamma_i x_i\}$$

1.2 Distances, normes et espaces

Définition 1.3 *Une fonction $f : \mathcal{X} \to \mathcal{X}$ est dite symétrique si*

$$f(x, y) = f(y, x), \forall (x, y) \in \mathcal{X} \times \mathcal{X} \tag{1.1}$$

Définition 1.4 *Une* **distance** *sur $\mathcal{V}$ est une fonction*

$$\mathrm{d}(.,.) : \mathcal{V} \times \mathcal{V} \to \mathbb{R} \tag{1.2}$$

$$(\mathbf{v}, \mathbf{v}') \mapsto \mathrm{d}(\mathbf{v}, \mathbf{v}') \tag{1.3}$$

telle que

(i) $\mathrm{d}(\mathbf{u}, \mathbf{u}) = 0$,

(ii) $\mathbf{u} \neq \mathbf{v} \Rightarrow \mathrm{d}(\mathbf{u}, \mathbf{v}) > 0$,

(iii) $\mathrm{d}(\mathbf{u}, \mathbf{v}) = \mathrm{d}(\mathbf{v}, \mathbf{u})$, *c.-à-d. que* d *est symétrique,*

(iv) $\mathrm{d}(\mathbf{u}, \mathbf{v}) \leq \mathrm{d}(\mathbf{u}, \mathbf{w}) + b(\mathbf{w}, \mathbf{v}), \forall \mathbf{w} \in \mathcal{V}$.

Définition 1.5 *Une* **norme** *est une fonction $\|.\| : \mathcal{V} \to \mathbb{R}^+$ telle que pour tous $x, x' \in \mathcal{V}$ et $\alpha \in \mathbb{K}$*

$$\|x + x'\| \leq \|x\| + \|x'\| \tag{1.4}$$

$$\|\alpha x\| = |\alpha| \, \|x\| \tag{1.5}$$

$$x \neq \mathbf{0} \Rightarrow \|x\| > 0. \tag{1.6}$$

Remarquons que toute norme définit implicitement une distance $\mathrm{d}(x, x') = \|x - x'\|$.

Définition 1.6 *Une* **forme bilinéaire** *sur un espace vectoriel $\mathcal{V}$ est une fonction*

$$b : \mathcal{V} \times \mathcal{V} \to \mathbb{R} \tag{1.7}$$

qui satisfait les trois axiomes suivants pour tout choix de scalaire $\alpha \in \mathbb{K}$ *et de vecteurs* $x, y, z \in \mathcal{V}$.

(i) $b(\alpha x, y) = b(x, \alpha y) = \alpha b(x, y)$

(ii) $b(x + y, z) = b(x, z) + b(y, z)$

(iii) $b(x, y + z) = b(x, y) + b(x, z)$

Définition 1.7 *Un* **produit scalaire** *sur un espace vectoriel* $\mathcal{V}$

$$\langle ., . \rangle_{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \to \mathbb{R} \tag{1.8}$$

$$(x, x') \mapsto \langle x, x' \rangle_{\mathcal{V}} \tag{1.9}$$

est une forme bilinéaire symétrique définie strictement positive sur $\mathcal{V}$ *c'est à dire que*

$$\langle x, x' \rangle_{\mathcal{V}} = \langle x', x \rangle_{\mathcal{V}} \tag{1.10}$$

et

$$\langle x, x \rangle_{\mathcal{V}} \geq 0, \forall x \in \mathcal{V} \tag{1.11}$$

$$\langle x, x \rangle_{\mathcal{V}} = 0 \Rightarrow x = \mathbf{0}. \tag{1.12}$$

L'indice spécifiant l'espace sera omis lorsque évident selon contexte. Soulignons aussi que sur l'espace euclidien $\mathbb{R}^n$, le produit scalaire usuel est

$$\langle x, x' \rangle = \sum_{i=0}^{n} [x]_i [x']_i \,. \tag{1.13}$$

Au cours de ce mémoire, sauf en cas de mention contraire, nous entendrons par produit scalaire sur $\mathbb{R}^n$ le produit scalaire usuel.

Définition 1.8

(i) Un **espace normé** *est un espace vectoriel muni d'une norme.*

(ii) *Un* **espace pré-hilbertien** *est un espace vectoriel muni d'un produit scalaire.*

(iii) *Un* **espace métrique** *est un espace vectoriel muni d'une distance.*

Les concepts d'espace normé et pré-hilbertien sont fortement reliés vu le fait qu'un produit scalaire définit implicitement une norme $\|x\| = \sqrt{\langle x, x\rangle}$. D'autre part, au cours de ce mémoire nous ne traiterons que de distances du type $\mathrm{d}(x, x') = \|x - x'\|$. Les espaces métriques que nous considérerons seront donc tous normés.

Proposition 1.2.1 (Cauchy-Schwartz) *Soit $\mathcal{V}$ un espace pré-hilbertien et $\mathbf{v}, \mathbf{v}' \in \mathcal{V}$, alors*

$$|\langle \mathbf{v}, \mathbf{v}\rangle| \leq \|\mathbf{v}\| \, \|\mathbf{v}'\| \, .$$

1.3 Suites et convergence

Nous introduisons d'entrée de jeu les notions de monotonie. Le lecteur sera à même d'apprécier que la prochaine définition clarifie ce que l'auteur entend lorsqu'il parle, par exemple, de croissance plutôt que de non décroissance.

Définition 1.9 *Une suite $\{x_n\}$ de nombres réels est dite*

(i) *croissante si $x_{n+1} \geq x_n, \forall n \in \mathbb{N}$;*

(ii) *strictement croissante si $x_{n+1} > x_n, \forall n \in \mathbb{N}$;*

(iii) *décroissante si $x_{n+1} \leq x_n, \forall n \in \mathbb{N}$;*

(iv) *strictement décroissante si $x_{n+1} < x_n, \forall n \in \mathbb{N}$.*

On dit **(strictement) monotone** *une suite qui respecte l'une de ces propriétés.*

Définition 1.10 *Soit $\mathcal{V}$ un espace métrique et $\{x_n\}$ une suite sur cet espace. On dit que $\{x_n\}$* **converge vers** *x si*

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ tel que } n > N \Rightarrow \mathrm{d}(x_n, x) < \epsilon.$$

La suite $\{x_n\}$ est alors dite **convergente**.

Définition 1.11 *Soit $\mathcal{V}$ un espace métrique et $\{x_n\}$ une suite sur cet espace. On dit que $\{x_n\}$ est une* **suite de Cauchy** *si*

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ tel que } n, m > N \Rightarrow \mathrm{d}(x_n, x_m) < \epsilon.$$

Définition 1.12

(i) Un espace $\mathcal{V}$ est dit **complet** *si toute suite de Cauchy sur l'espace converge.*

(ii) Un **espace de Banach** *est un espace normé complet.*

(iii) Un **espace de Hilbert** *est un espace pré-hilbertien complet.*

Ayant défini les espaces de Hilbert, nous pouvons introduire une définition simple des projections d'un espace à un de ses sous-espaces

Définition 1.13 *Soient $\mathcal{H}$ un espace de Hilbert, E une partie convexe fermée non vide de $\mathcal{H}$. Alors la fonction*

$$\mathcal{H} \to E \tag{1.14}$$

$$x \mapsto \mathrm{proj}_E(x) := argmin_{y \in E} \|y - x\|_{\mathcal{H}} \tag{1.15}$$

est dite la **projection de** *x* **dans** *E.*

Maintenant, il nous sera profitable de nous remémorer quelques propriétés des suites de Cauchy quant à la convergence.

Théorème 1.1

(i) Dans un espace métrique, toute suite convergente est de Cauchy.

(ii) Toute suite convergente d'un espace métrique compact $\mathcal{V}$ converge dans $\mathcal{V}$.

(iii) Toute suite de Cauchy sur $\mathbb{R}^n$ est convergente. L'espace euclidien muni de son produit scalaire usuel est donc de Hilbert.

Définition 1.14 *L'espace l_p^n est défini comme étant l'espace vectoriel $\mathbb{R}^n$ muni de la norme p :*

$$1 \leq p \leq \infty : \|x\|_{l_p^n} := \|x\|_p = \left(\sum_{i=1}^{n} |[x]_i|^p \right)^{1/p}$$
$$p \to \infty : \|x\|_{l_\infty^n} := \|x\|_\infty = \max_{i=1,\ldots,n} |[x]_i|.$$

Il est d'usage de noter $l_p := l_p^{n \to \infty}$, où l'on ne considère évidemment que les suites de norme p finie. Aussi, le maximum devient-il un supremum dans $\|x\|_{l_\infty^n}$.

Notons que, étant donné une fonction $f \in \mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$, on peut définir la norme p de f sur un échantillon $\mathrm{S} = \{x_1, \ldots, x_m\}$ comme étant

$$1 \leq p \leq \infty : \|f\|_{l_p^{\mathrm{S}}} = \|(f(x_1), \ldots, f(x_m))\|_{l_p^m}$$
$$p \to \infty : \|f\|_{l_\infty^{\mathrm{S}}} = \max_{i=1,\ldots,m} |f(x_i)|.$$

Quoique utile, l'extension précédente n'est pas suffisante à l'analyse de fonction indépendemment de quelque ensemble de données. C'est pourquoi la définition de norme p sur une fonction $f \in \mathcal{F}$ sera, étant donné une mesure μ sur une

sigma-algèbre donnée*,

$$1 \leq p \leq \infty : \|f\|_{L_p^{\mathcal{X}}} := \|f\|_p = \left(\int |f(x)|^p d\mu(x) \right)^{1/p}$$
$$p \to \infty : \|f\|_{L_\infty^{\mathcal{X}}} := \|f\|_\infty \operatorname*{ess\,sup}_{x \in \mathcal{X}} |f(x)|$$

où $\operatorname{ess\,sup}_y g(y)$ est **le suprémum essentiel** de g en y au sens où l'ensemble des points pour lesquels $g(y) > \operatorname{ess\,sup}_y g(y)$ est de mesure nul.

Définition 1.15 *On dénome l'***espace** $L_p^{\mathcal{X}}$ *l'espace de fonctions sur* $\mathcal{X}$ *défini par*

$$L_p^{\mathcal{X}} := \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_p < \infty\}.$$

1.4 Quelques notions d'algèbre

Soit M, une **matrice réelle** $n \times m$. Nous dénoterons par $[\mathrm{M}]_{i,j}$ l'élément de M qui se trouve à l'intersection de sa i^e ligne et de sa j^e colonne. Conséquemment, les notations $[\mathrm{M}]_{i,.}$ et $[\mathrm{M}]_{.,j}$ désignerons respectivement la i^e ligne et la j^e colonne de M, que nous considérerons comme étant des vecteurs.

La matrice M^{T} est dite la **transposée** de M et est définie telle que $[\mathrm{M}]_{i,.} = [\mathrm{M}]_{.,i}{}^{\mathrm{T}}$. On dit **carrée** une matrice telle que $n = m$ et **symétrique** une matrice carrée telle que $\mathrm{M} = \mathrm{M}^{\mathrm{T}}$.

*Les concepts de **mesure** et de **sigma-algèbre**, pour être bien introduits, nécessiteraient la tenu d'un discours mathématique auquel le présent mémoire ne prétend pas. Pour une introduction rigoureuse aux notions de mesure et intégration, le lecteur peut se référer à (BILLINGSLEY 1995).

Définition 1.16 *Une matrice $n \times n$ M est dite* **définie positive** *si pour tout vecteur $x \in \mathbb{R}^n, x \neq 0$,*

$$x^T \mathrm{M} x \geq 0. \tag{1.16}$$

Nous dirons M définie **strictement**[†] *positive si l'inégalité précédente est stricte.*

Une matrice M est dite **définie (strictement) négative** *si $-\mathrm{M}$ est définie (strictement) positive.*

Au cours de ce mémoire, nous considérerons aussi comme acquises les notions relatives au vecteurs et valeurs propres d'une matrice.

Définition 1.17 *On dit $\lambda_i \in \mathbb{C}$, où $\mathbb{C}$ dénote l'ensemble des nombres complexes, valeur propre d'une matrice $n \times n$ M et $v_i \in \mathbb{C}^n, v_i \neq 0$, le vecteur propre associé si le couple (λ_i, v_i) respecte*

$$\mathrm{M} v_i = \lambda_i v_i. \tag{1.17}$$

M compte n valeurs propres, possiblements répétées, et $\{\lambda_i\}_{i=1}^n$ est souvent dit le **spectre** *de M.*

Proposition 1.4.1 *Soit M une matrice $n \times n$ symétrique. Alors :*

(i) Les valeurs propres λ_i de M sont toutes réelles.

(ii) Les vecteurs propres v_i de M sont à composante réelles et forme une base de $\mathbb{R}^n$.

(iii) Supposons que l'on ait normalisé les v_i ($\|v_i\| = 1, \forall i$), alors

$$\mathrm{M} = \sum_{i=1}^{n} \lambda_i v_i v_i^T. \tag{1.18}$$

[†] *Plusieurs auteurs préférerons définir comme semi définie positive une matrice entraînant l'inégalité et comme définie positive une matrice entraînant l'inégalité stricte. Quoique moins répandue, la notation introduite ici a l'avantage d'être plus cohérente avec la terminologie des noyaux définis positifs (Définition (7.3)).*

De même, certains résultats nécessiteront un recours aux notions d'opérateur et de fonctions propres d'un opérateur.

Définition 1.18 *Soit $\mathcal{H}$ un ensemble de fonctions. Une transformation de fonctions dans $\mathcal{H}$*

$$O : \mathcal{H} \to \mathcal{H} \tag{1.19}$$

$$f \mapsto g \tag{1.20}$$

est un **opérateur**.

On distingue, entre autres, les **opérateurs linéaires**. *Un opérateur est dit linéaire s'il satisfait*

$$O[\lambda_1 f_1 + \lambda_2 f_2] = \lambda_1 O[f_1] + \lambda_2 O[f_2]. \tag{1.21}$$

Définition 1.19 *Soit O un opérateur sur $\mathcal{H}$. Une fonction $f \in \mathcal{H}$ telle que*

$$O[f] = \lambda f \tag{1.22}$$

est une **fonction propre** *de O avec λ la* **valeur propre** *associée.*

Partie II

Le monde de la finance

Pour les non initiés, les pages boursières des quotidiens ne sont rien de plus qu'un raz-de-marée de petits chiffres, interface incompréhensible d'une jungle effervescente. Toutefois, malgré le chaos apparent, les marchés financiers sont des milieux les plus organisés qui soient. La grande majorité des acteurs principaux du milieu s'inspire d'une théorie rigoureusement établie et largement documentée. Ajoutez une réglementation des plus strictes et il devient évident que tout investisseur qui se veut sérieux se doit de s'instruire un minimum sur la mécanique du monde de la finance.

Nous n'aurions pas la prétention de pouvoir, en quelques courts chapitres, faire le tour de la question, mais nous espérons faire un survol[a] qui puisse donner une intuition suffisante au lecteur pour saisir le contexte d'application de nos méthodes.

[a] L'organisation de ce chapitre est inspirée de l'évolution du livre *Investments, Fourth Canadian Edition* (Bodie, Kane et Marcus 2003) que nous suggérons à quiconque désire une introduction rigoureuse et complète aux enjeux sous-jacents à l'investissement.

CHAPITRE 2

Les instruments financiers

Pour commencer cette partie, nous irons d'une petite introduction sur les différents biens pouvant être considérés dans la gestion d'un portefeuille. Il ne s'agira que d'un survol dont le but véritable est en fait d'introduire proprement au lecteur les contrats à terme boursiers, biens considérés dans l'évaluation du modèle suggéré par ce mémoire.*

2.1 Les instruments de dette

À la base de l'investissement : la recherche d'un équilibre entre les besoins actuels et futurs. En effet, une disponibilité de capitaux chez les uns et un besoin en capital chez les autres ont vite fait de donner lieu au mécanisme fondamental de l'investissement : le prêt. Un simple dépôt à la banque se veut un prêt — à un taux dérisoirement bas ? — d'un individu à l'institution financière qui encaisse le dépôt. Pourquoi le taux auquel un individu "prête" à sa banque est-il tellement plus bas que celui auquel une banque prête à un

*Pour une couverture plus exhaustive des marchés et des instruments financiers le lecteur peut consulter, bien évidemment, (BODIE, KANE et MARCUS 2003) ainsi que (STIGUM 1989) et (LOGUE 1994) pour les marchés américains ainsi que (SARPKAYA 1989) et (SMITH et AMOAKA-ADU 1990) pour les marchés canadiens.

particulier ? Tout est une question de compromis entre le **rendement espéré** et le **risque** encouru. Nous reviendrons à cet enjeu au chapitre 3.

Ainsi, l'instrument financier fondamental est le bon d'épargne. Périodiquement, le gouvernement canadien émet des bons d'épargne, appelés bons du Trésor, par le biais desquels l'investisseur se trouve à prêter une somme donnée au gouvernement en échange d'un certain taux d'intérêt. Les bons du Trésor, qui viennent à échéance après six mois, sont considérés virtuellement sans risque car, compte tenu de son pouvoir de taxation, le gouvernement devrait toujours être en mesure de payer ses dûs. Les obligations d'épargne, quant à elles, sont de plus longue durée. C'est pourquoi leur taux d'intérêt est habituellement plus élevé, car l'investisseur court un risque plus grand face à l'inflation (voir (§ 3.1.2)).

Ces instruments de dette existent aussi pour les paliers de gouvernement provinciaux et municipaux. Or moins le pouvoir de taxation est grand, plus le taux d'intérêt doit être élevé pour attirer les investisseurs. De même, les entreprises peuvent recourir à des instruments semblables, mais encore une fois, la possibilité de faillite se reflète dans les taux d'intérêt qu'elles doivent offrir pour concurrencer les instruments plus sûrs.

2.2 Les actions

Associées à l'image d'un crieur aux nerfs d'acier qui agit sur le parquet de Wall Street, véritables symboles des marchés financiers, les actions sont sans nul doute l'objet financier le plus connu du grand public. Que sont-elles en fait ?

L'action de base[†] procure les droits, à celui qui la possède, sur une partie de la propriété d'une compagnie. Chaque action procure à son propriétaire un vote à

[†]Il existe plusieurs autres types d'actions. Le principe de base est toujours semblable, mais elles ne procurent pas toutes les mêmes droits.

l'assemblée annuelle des actionnaires, où seront soumises les décisions majeures des dirigeants de l'entreprise, et assure à l'actionnaire le droit de toucher une partie des profits de la compagnie, remis sous la forme de **dividendes**. Il est donc important de remarquer que la possession d'une action est associée à la possession *véritable*, quoique partielle, d'une compagnie. Ainsi, le prix d'une action devrait correspondre, *grosso modo*, à la valeur totale de toute propriété de la compagnie, autant matérielle qu'intellectuelle, divisée par le nombre d'actions en circulation.

2.3 Les options

Des nombreux marchés qui meublent le paysage de la finance moderne, le marché des produits dérivés[‡] est sans conteste celui qui a connu le plus grand essor dans les dernières années. Les produits dérivés doivent leur nom au fait que leur valeur n'est pas directement reliée à une propriété véritable, mais plutôt à la valeur d'autres biens, que nous dirons **sous-jacents**. De ce groupe, nous introduisons ici les options, puis, à la section suivante, les contrats à terme boursiers.

Une option, comme sont nom l'indique, vaut à son propriétaire un droit, mais non une obligation, d'acheter ou de vendre dans le futur, suivant qu'il s'agisse d'une **option d'achat** ou d'une **option de vente**, un sous-jacent à un prix fixé d'avance, le **prix d'exercice**. La **maturité** d'une option est le temps T auquel elle vient à échéance.

Il y a plusieurs types d'options, les deux plus connus étant les options américaines et les options européennes. La différence fondamentale entre les deux types d'options est que, pour les options américaines, le détenteur de l'option peut se prévaloir de son droit d'achat ou de vente sur toute la période allant

[‡]Pour un exposé complet et rigoureux sur les produit dérivé, un excellent livre est *Options, Futures and Other Derivatives* (HULL 2003).

de l'achat de l'option à la maturité. Par contre, le détenteur d'une option européenne ne peut se prévaloir de son droit qu'au jour de la maturité. En dénotant le prix d'exercice par K et le prix du sous-jacent au temps t par S_t on observe

Valeur d'une option d'achat européenne

$$V = \begin{cases} S_T - K & : \quad S_T > K \\ 0 & : \quad S_T \leq K \end{cases}$$

Valeur d'une option de vente européenne

$$V = \begin{cases} K - S_T & : \quad K > S_T \\ 0 & : \quad K \geq S_T \end{cases}$$

Un lecteur averti se sera peut-être déjà dit que la valeur d'une option américaine est plus difficile à déterminer, puisqu'elle dépend de la politique de l'investisseur quant à l'exercice de son droit d'achat ou de vente. Il est intéressant de remarquer que la valeur d'une option est une variable aléatoire toujours non négative. Le rendement $(V - p)$ quant à lui dépendra de la **prime** p défrayée par l'investisseur pour acquérir l'option. Une fois combinées, plusieurs options de types et de maturités différentes peuvent donner lieu à de très élégantes stratégies d'assurance ou de spéculation. C'est pourquoi les options sont un des outils favoris des ingénieurs financiers, qui se sont d'ailleurs dotés d'une vaste gamme d'options exotiques dont les fonctions de valeur, plus complexes, permettent de développer des stratégie d'investissement d'autant plus subtiles.

2.4 Les contrats à terme boursiers

Le principe des **contrats à terme**[§] ne date pas d'hier. Il suffit que deux parties fixent un prix et s'engagent d'une part à vendre et d'autre part à acheter un bien pour ce prix dans un futur plus ou moins éloigné. Contrairement à une option, un contrat à terme entraîne une obligation de vente ou d'achat. Un producteur de blé, par exemple, peut ainsi convenir d'avance avec une

[§]En anglais : *forward contracts*

boulangerie d'une quantité de blé à transiger à la fin de la récolte et d'un prix auquel la transaction sera conclue. Ainsi, s'il s'avérait que la saison soit mauvaise, le boulanger se serait protégé contre une possible hausse des prix imputable à une offre amoindrie. À l'inverse, si la récolte était particulièrement bonne cette saison là, le producteur serait à l'abri d'une éventuelle baisse des prix.

Simple formalisation du concept, les **contrats à terme boursiers*** permettent de transiger de tel contrats de façon plus systématique. Les modalités des contrats sont fixées d'avance. Ainsi, il ne reste qu'à convenir d'un prix, qui évoluera, en fait, selon l'offre et la demande sur le marché†. Un des avantages majeurs de cette formalisation est qu'elle permet une plus grande liquidité‡ des contrats, le marché mettant en contact plusieurs investisseurs et permettant à chacun de se départir de ses obligations relatives à un contrat en les échangeant à un autre investisseur, et cela sans avoir à échanger concrètement le bien sous-jacent.

2.4.1 Terminologie

Nous noterons la valeur à laquelle on estime un contrat lorsqu'il est mis en marché, nous dirons **ouvert**, par F_0 et la valeur à la t^e période ultérieure par F_t. Comme pour les options, nous noterons par T la période correspondant à la maturité du contrat. Il est à remarquer que, tout naturellement, plus l'échéance d'un contrat approche, plus sa valeur se rapproche de la valeur du sous-jacent sur le marché, de telle façon que F_T est exactement la valeur du sous-jacent. Cette réalité est appelée **propriété de convergence**.

*En anglais : *future contracts* ou simplement *futures*

†Les contrats considérés dans ce mémoire sont échangés sur le *Chicago Board Of Trade*.

‡Liquidité : L'aisance avec laquelle on peut convertir un bien en argent (BODIE, KANE et MARCUS 2003). Un investissement dans l'immobilier, par exemple, offre moins de liquidité que l'achat d'actions concurrentielles, car il est plus facile de revendre rapidement ces dernières que de se départir d'un bâtiment.

Nous dirons que celui qui s'engage à acheter le bien sous-jacent prend une **position longue** et que celui qui s'engage à vendre prend une **position courte**. Par abus de langage, on dira souvent que celui qui prend la position longue achète le contrat et que celui qui prend la position courte vend le contrat, bien qu'aucune somme d'argent ne soit transigée entre les deux parties qui ne font en fait que s'entendre sur un contrat. Aussi, dira-t-on parfois qu'un investisseur qui sort de la position longue (ref. courte) *vend* (ref. *achète*) son contrat à celui qui prendra cette position à sa place[§]. Ces abus de langage permettent un discours plus efficace et ne sont pas nuisibles dans la mesure où l'on garde en mémoire les précédentes nuances.

2.4.2 Les mécanismes d'échanges

Il est fondamental de savoir qu'au centre de tout échange des contrats à terme boursiers se trouve la **chambre de compensation**. Effectivement, sur les marchés de contrats à terme boursiers, le tenant de la position longue et celui de la position courte ne font jamais affaire ensemble. En fait, lorsqu'un investisseur est intéressé à prendre une position longue, par exemple, c'est la chambre de compensation qui prend la position courte associée. Or, la chambre ne prend cette position que si elle trouve, d'autre part, un investisseur prêt à prendre une position courte équivalente dans un contrat où la chambre tiendra cette fois la position longue. La position nette de la chambre est donc toujours nulle.

À première vue, cet artifice peut sembler alourdir les échanges, mais, au contraire, cette institution est le noyau de l'efficacité des marchés de contrats à terme boursiers, mettant en vigueur une vaste gamme de mécanismes dont nous entamons ici un survol[¶].

[§]En fait, on appelle **échange inversé** l'échange visant à se défaire des obligations relatives à un contrat. Ce mécanisme est pratiquement omniprésent, comme nous le verrons ultérieurement ((§ 2.4.2)).

[¶]Pour une couverture exhaustive des différents mécanismes, voir (Bodie, Kane et Marcus 2003; Hore 1987; Khoury et Laroche 1995; Hull 2003).

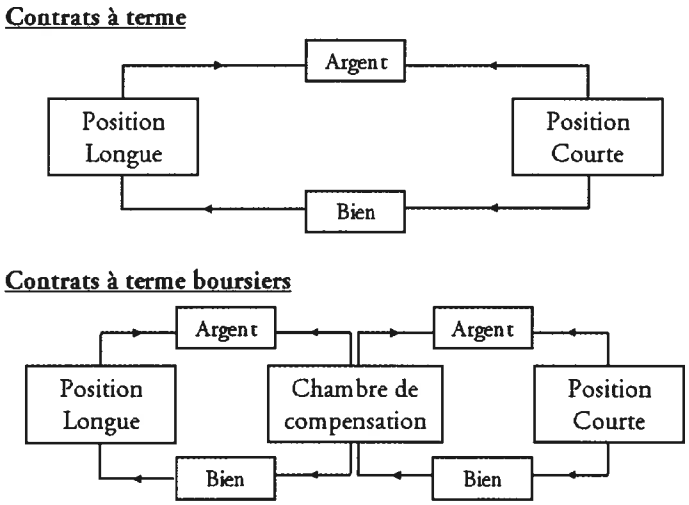

Figure 2.1 – *La Chambre de compensation*

D'abord et avant tout, advenant que l'une ou l'autre des parties faille à ses obligations, la chambre serait la seule à en subir les conséquences. Ainsi, un investisseur obtient l'assurance que le contrat dans lequel il s'engage sera respecté. D'autre part, la position institutionnelle de la chambre lui confère des pouvoirs légaux lui permettant de veiller à ce que chacun respecte ses engagements beaucoup plus aisément que ne pourrait le fait n'importe quel investisseur individuel.

Aussi, la présence de la chambre permet-elle de liquider sa position avec beaucoup plus de facilité. Advenant qu'un investisseur ait une position longue dans un contrat donné et qu'il désire liquider cette position, il lui suffit d'entrer dans une position courte sur le même contrat et les positions s'annuleront une fois que la chambre aura trouvé un investisseur prêt à entrer dans la position longue. Ce mécanisme, nommé **échange inversé**[‖], est omniprésent sur les marchés de contrats à terme boursiers. Lorsqu'une banque, par exemple, entre dans une position longue sur un contrat l'engageant à acheter une tonne de blé, il est fort peu probable qu'elle soit effectivement intéressée à éventuellement

[‖]Traduction libre de l'anglais *reversing trade*

recevoir ce blé. Ainsi, à la veille de l'échéance du contrat, il lui suffit d'entrer dans un échange inversé pour bénéficier des fruits éventuels de son contrat sans n'avoir jamais reçu de blé. La plupart des investisseurs qui transigent sur ces marchés ne sauraient ainsi que faire des biens sous-jacents aux contrats, ce qui explique que la grande majorité des contrats se terminent dans un échange inversé**.

L'**indice de positions ouvertes** est le nombre de contrats d'un type donné actuellement en circulation. Il est donc égal au nombre de positions longues (ou courtes) prises par la chambre de compensation. L'indice est nul lors de l'ouverture du contrat puis augmente avec le nombre de contrats dans lesquels s'engagent les investisseurs. À l'approche de la maturité, l'indice sur ce type de contrat diminue au fur et à mesure que les investisseurs entrent dans des échanges inversés.

Marge et valorisation au prix du marché

Comme un investisseur n'achète ni ne vend aucun bien au moment d'entrer dans une position donnée, il pourrait bien n'y avoir aucune transaction monétaire impliquée. Or, une fois l'échéance venue la chambre de compensation pourrait avoir la mauvaise surprise de constater l'impossibilité pour l'investisseur de remplir ses obligations. Ainsi, pour éviter tout problème, chaque fois qu'un investisseur entre dans une nouvelle position, la chambre exige de l'investisseur un dépôt de garantie dans un compte appelé la **marge**. La marge est habituellement investie au taux sans risque, l'investisseur tirant donc tout de même un rendement de l'argent qu'il y verse (voir (2.2)).

Dépendemment de sa confiance dans l'investisseur, la chambre peut permettre à ce dernier de déposer un montant inférieur à la valeur du contrat. Le facteur

** "The image of a trader awakening one delivery date with a hog in the frontyard is amusing, but unlikely." (BODIE, KANE et MARCUS 2003)

suivant lequel la relation entre le dépôt et la valeur est établie est appelé le **levier**[††].

Définition 2.1 *Soient F_t la valeur d'un contrat et m_t la marge d'un investisseur au temps t. Si n_t est le nombre de contrats de ce type dans lesquels cet investisseur est impliqué, on dit alors qu'il jouit d'un* **levier** *l où*

$$l = \frac{n_t F_t}{m_t}. \tag{2.1}$$

Exemple 2.1 La Marge
Un investisseur désire prendre une position longue dans 3 contrats dont le sous-jacent est le blé. La valeur actuelle du contrat est de 19 075\$. Si la chambre accorde à cet investisseur un levier de 10, l'investisseur devra déposer à sa marge 5722.50\$ $(\frac{3*19075}{10})$, *soit 10% de la valeur actuelle des contrats.*

◇

Remarquons que le levier accordé à un même investisseur peut être différent d'un contrat à l'autre suivant la volatilité des prix des sous-jacents.

Sous la notation introduite précédemment, le gain réalisé par le tenant d'une position longue (resp. position courte) qui liquide cette dernière au temps $\bar{t}$ sont de $F_{\bar{t}} - F_0$ (resp. $F_0 - F_{\bar{t}}$). Or, plutôt que d'attendre que l'investisseur liquide sa position pour lui charger ou lui remettre ce qui lui est dû, la chambre dépose ou retire de la marge la variation à chaque période t, $F_t - F_{t-1}$ (resp. $F_{t-1} - F_t$)*. Ce procédé est appelé **valorisation au prix du marché**[†].

Définition 2.2 *Soient r_f le taux sans risque en vigueur entre $t-1$ et t, F_t la valeur d'un contrat et m_t la marge d'un investisseur au temps t. Si n_t est*

[††]Les investisseurs qui utilisent les contrats à terme sont pour la plupart des spéculateurs. Un des grands avantages des contrats à terme est d'offrir à l'investisseur l'occasion d'obtenir des rendements élevés sur des placements relativement modestes : c'est ce qu'on appelle le levier financier. Le revers de la médaille, c'est que ce levier financier peut aussi se traduire par des pertes considérables s'il est mal utilisé.

*Évidemment, $F_{\bar{t}} - F_0 = \sum_{t=1}^{\bar{t}} F_t - F_{t-1}$ (resp. $F_0 - F_{\bar{t}} = \sum_{t=1}^{\bar{t}} F_{t-1} - F_t$).

[†]De l'anglais : *Marking the market*

le nombre de contrats de ce type dans lesquels cet investisseur est impliqué, la **valorisation au prix du marché** *est le mécanisme selon lequel*

$$m_{t+1} = (1 + r_{tf})m_t + n_t(F_t - F_{t-1}). \tag{2.2}$$

Rappelons que le facteur $(1+r_f)$ *est dû à l'investissement de la marge au taux sans risque*[‡].

Ce faisant, le montant inscrit à la marge peut diminuer nettement sous le seuil initialement exigé par le levier. La chambre fixe donc une valeur critique, la **marge de maintenance**, sous laquelle la marge ne doit pas tomber. Advenant que la situation se présente, l'investisseur reçoit un **appel de marge** : soit il renfloue la marge, soit la chambre liquidera suffisamment de positions pour que la valeur de la marge suffise.

Exemple 2.2 La Marge (suite)
Notre investisseur a une marge de maintenance de 7,5%. Or, en quelques jours, le prix du blé a dramatiquement chuté. Sous l'effet de cette baisse, la valeur d'un contrat sur le blé est tombée à 18275$. Ainsi, la marge de notre investisseur n'est plus que de $3322.50\$$ $(5722.50 - 3 * (19075 - 18275))$, *soit 6.05% des* $54825\$$ *que valent les trois contrats. Il reçoit donc un appel de marge. Compte tenu de son levier de 10, l'investisseur doit déposer* $2160\$$ $(\frac{3*18275}{10} - 3322.5)$ *sinon la chambre liquidera 2 de ses contrats, portant sa marge au taux de 18,1% de la valeur de la position restante.*

◇

[‡]Notons ici que le facteur $(1 + r_{tf})$ doit être établi à l'aide du taux d'intérêt en vigueur et *pour* la période t. Le taux d'intérêt publié est pratiquement toujours un taux d'intérêt annualisé, c'est-à-dire qu'il représente un taux composé en fonction du nombre de périodes dans l'année. Pour obtenir un taux quotidien r_f^{quot}, par exemple, à partir d'un taux annualisé r_f^{an} il faut appliqué, pour un taux annualisé en continue (BODIE, KANE et MARCUS 2003), la formule

$$r_f^{\text{quot}} = \frac{\exp(\log(1 + r_f^{\text{an}}))}{d} - 1, \tag{2.3}$$

où d est le nombre de jours ouvrables dans l'année.

CHAPITRE 3

Théorie moderne du portefeuille

Investir est un procédé qui consiste essentiellement en deux étapes. Dans un premier temps, une analyse des marchés et des biens en présence s'impose. Ensuite, il faut construire un **portefeuille** que l'on voudrait idéal en regard de la précédente analyse. Par portefeuille, on entends une caractérisation de la répartition du capital dont l'individu dispose entre les différentes options d'investissements qu'il envisage. Cette tâche de répartition relève de la **théorie du portefeuille**.

3.1 Concepts sous-jacents

3.1.1 Rendements

À la base de tout investissement, le désir de faire fructifier des capitaux. La notion de rendement, mesure de cette fructification, sera donc omniprésente tout au long de ce mémoire.

Définition 3.1 *Les rendements R_t sur un bien donné sont des variables aléatoires de densités μ_t inconnues. Nous noterons r_t une réalisation de R_t. Notons que*

(i) L'indice f (R_{ft}, r_{ft}) est strictement reservé à l'usage du taux sans risque (risk free rate).

(ii) Les rendements sur des biens risqués seront indexés d'entiers $k \in \mathbb{N}$ (R_{kt}, r_{kt}). Toutefois, cet index sera parfois omis si un seul bien est considéré.

(iii) Les indices majuscules (ex. R_{Pt}, r_{Pt}) seront strictement reservés pour traiter du rendement de portfeuilles composés de plusieurs actifs.

(iv) L'indice de temps peut être négligé (ex. r_f) s'il est clair qu'on ne traite que d'une période arbitrairement choisie.

3.1.2 Taux d'intérêt

Comme le concept de risque est central à la théorie moderne du portefeuille, il est naturel de s'attarder en premier lieu sur ce que serait un **bien sans risque**. Normalement, les taux d'intérêt émis sur les bons du Trésor par la Banque du Canada sont considérés comme étant virtuellement sans risque. En effet, lorsqu'un investisseur achète un bon du Trésor, il prête au gouvernement canadien la somme pour laquelle il achète le bon. Vue la forte improbabilité d'une faillite du gouvernement, on considère l'investissement sans risque.

Toutefois, une distinction entre les **taux d'intérêt nominaux** et les **taux d'intérêt réels** s'impose. Les taux qui sont publiés dans les cahiers boursiers sont les taux nominaux tels que fixés par la Banque du Canada compte tenu

(i) des fonds dont les ménages disposent pour l'épargne ;

(ii) de la demande en capitaux des entreprises pour l'achat de biens tangibles et la formation d'un capital humain ;

(iii) de l'offre et/ou de la demande de fonds nette du gouvernement telles qu'établies par les autorités monétaires du pays.

Bien que souvent nous considérerons le taux sans risque comme étant le taux nominal r_{nominal} sur les bons du Trésor, il faut toujours se rappeler que les taux nominaux demeurent exposés à un type de risque : l'inflation. Il faut

en effet avoir conscience que la croissance réel $r_{\text{réel}}$ du pouvoir d'achat d'un investisseur augmente avec l'augmentation de son avoir divisée par le nouveau niveau des prix compte tenu d'une inflation i, c'est-à-dire

$$1 + r_{\text{réel}} = \frac{1 + r_{\text{nominal}}}{1 + i} \tag{3.1}$$

$$\Leftrightarrow$$

$$r_{\text{réel}} = \frac{r_{\text{nominal}} - i}{1 + i} \tag{3.2}$$

Le taux sans risque réel r_f est donc fonction du taux d'inflation sur la période de prêt, qui n'est évidemment pas connu avant la fin de la période. Voilà pourquoi nous considérerons le taux sans risque r_f comme étant le taux nominal, en supposant que le taux d'inflation *sur une courte période* est négligeable.

Définition 3.2 *Nous considérons que le taux sans risque r_{ft} est égal à $r_{nominal,t}$, c'est-à-dire que*

$$\Pr(R_{ft} = r_{nominal,t}) = 1, \forall t. \tag{3.3}$$

Rappelons que, tout au long de ce mémoire, l'indice f est réservé au rendement sur le bien sans risque. Aussi doit-on remarquer que l'ajout d'un indice de temps t à r_{ft} souligne que l'on considére le taux sans risque en vigueur sur la période t.

3.1.3 Risque et primes de risque

Le risque est un reflet de l'**incertitude** sur les rendements futurs d'un investissement. Par rendement, on entend une mesure de l'accroissement ou de la décroissance de la valeur d'un investissement. Par exemple, le taux d'intérêt sans risque est le **rendement attendu** d'un investissement dans les bons du Trésor. Un lecteur averti aura pressenti l'importance des probabilités et — à plus forte raison, nous le verrons — des statistiques dans le cadre d'une allocation des actifs soumise à la théorie du portefeuille. Rappelons que le

rendement R_t sur un bien risqué au temps t est une variable aléatoire de densité μ_t. Alors les concepts de rendement attendu et d'incertitude peuvent être associés à ceux d'espérance et de variance

$$\mathrm{E}[R_t] = \int r_t \, \mu_t(r_t) \, dr_t \tag{3.4}$$

$$\sigma_t^2 = \int (r_t - \mathrm{E}[R_t])^2 \, \mu_t(r_t) \, dr_t. \tag{3.5}$$

Quoique nous passions par la variance σ_t^2 pour calculer l'incertitude, nous considérerons la réelle mesure du risque comme étant l'écart-type σ_t, cette dernière mesure ayant l'avantage majeur d'être dans les mêmes unités que l'espérance.

Or, supposer que l'on connaisse les lois μ_t qui gouvernent l'évolution du processus des rendements est une présomption majeure. Ceci étant, la majorité des modèles se sont développés de façon à ne considérer que les deux premiers moments de ces densités : la moyenne et la variance. Ainsi, une approche statistique du problème peut nous permettre d'utiliser, au temps t, les informations contenues dans les rendements historiques r_τ, $\tau \leq t$, pour construire des estimateurs de ces deux paramètres :

$$S_{t,l} := \sum_{\tau=1}^{t} r_\tau^l$$

$$\overline{r}_t := \frac{S_{t,1}}{t} \tag{3.6}$$

$$\hat{\sigma}_t^2 := \frac{tS_{t,2} - S_{t,1}^2}{t(t-1)} \tag{3.7}$$

où $\overline{r}_t$ et $\hat{\sigma}_t^2$ sont, respectivement, des estimateurs au temps t de l'espérance et de la variance des rendements (Rice 1994).*

*L'estimateur du risque sera donc $\hat{\sigma}_t$.

Prime de risque

Quel que soit le moyen par lequel on évalue le rendement espéré sur un bien et son incertitude, le risque a un prix.

Définition 3.3 Prime de risque

(i) La différence $(r - r_f)$ *entre le rendement obtenu sur un bien risqué et le taux sans risque sur la même période est appelée le* **rendement excédentaire**.

(ii) La **prime de risque** *associée à un bien risqué est l'espérance du rendement excédentaire qu'il engendrera, c'est à dire*

$$\mathrm{E}[R] - r_f \tag{3.8}$$

3.1.4 Aversion au risque

Le concept de prime de risque, quoique fondamental, ne prend vraiment tout son sens que lorsque mis en rapport avec celui d'**aversion au risque**. Avant de formaliser cette nouvelle notion, abordons-la d'abord à l'aide d'un exemple simpliste.

Exemple 3.1 Investir ou non

Monsieur X viens de vendre sa maison et part pour une croisière de six mois au retour de laquelle il achètera une nouvelle demeure. Évidemment, il n'est pas question pour lui de laisser dormir ses 200 000$ dans un quelconque compte d'épargne. Il envisage d'investir cette somme dans une compagnie ABC dont l'action se vend présentement à 4$. Selon son conseiller financier, dans six mois l'action aura grimpé à 5$ si le contrat présenti avec CDE est effectivement signé. Toujours selon son conseiller, il demeure une probabilité de 0.4 que le contrat ne soit pas entériné par ABC, auquel cas l'action perdra de la valeur, chutant à 3.80$. Par ailleurs, sur la même période le taux sans risque est de 5%.

Pour prendre une décision éclairée, il faut dans un premier temps évaluer les rendement attendu et incertitude de l'investissement envisagé.

$$\begin{aligned}
\mathrm{E}[R] &= 0.6 * \frac{5.00 - 4.00}{4.00} + 0.4 * \frac{3.80 - 4.00}{4.00} \\
&= 0.6 * 0.25 + 0.4 * (-0.05) \\
&= 0.13 \\
\mathrm{Var}(R) &= 0.6 * (0.25 - 0.175)^2 + 0.25(-0.05 - 0.175)^2 \\
&= 0.0216
\end{aligned}$$

Le rendement attendu est de 13.0% avec un écart-type de 14.7%. En investissant la somme de 200 000$ maintenant dans XYZ, Monsieur X peut espérer récupérer environ 226 000$ dans six mois, mais ce avec un écart-type de 29 400$.

◊

Ainsi, la prime de risque dans notre exemple est $\mathrm{E}[R] - r_f = 0.08$, c'est-à-dire 8%. Quoique cette prime soit considérable, le lecteur aura sûrement compris qu'elle ne doit pas être considérée de façon absolue. En effet, comme la somme à investir est le résultat de toute une vie de travail et qu'au retour l'investisseur aimerais bien avoir un toît, il n'est pas nécessairement enclin à prendre de grands risques, on peut donc dire qu'il est **averse au risque**. Avant de prendre sa décision, il devra donc pénaliser le rendement espéré proportionnelement au risque encouru. Une formalisation classique de cette notion de pénalisation du risque est la fonction d'**utilité**. Plusieurs fonction d'utilité peuvent avoir leur justification dépendemment des contextes. En voici

une communément utilisée en gestion de portefeuille[†] :

$$\mathrm{U} = \mathrm{E}[R] - \frac{1}{2}\mathrm{A}\sigma^2 \tag{3.9}$$

où $\mathrm{A} > 0, \mathrm{A} \in \mathbb{R}$, est une mesure de l'aversion au risque propre à chaque investisseur[‡]. Plus A est élevé, plus l'investisseur est averse au risque. Remarquons que pour $\mathrm{A} > 0$ cette fonction est effectivement proportionnelle au rendement espéré et inversement proportionnelle au risque encouru.

En regard du risque existent aussi les concept d'individu neutre au risque ($A = 0$) ou même amateur de risque ($A < 0$). Quoiqu'il en soit, dans un contexte d'investissement on considère l'aversion au risque comme fondamentale et ce depuis l'avénement du paradoxe de Saint-Petersbourg, il y a plus de trois siècles et demi.

Exemple 3.2 Paradoxe de Saint-Petersbourg
De 1725 à 1738, Daniel Bernouilli, de la célèbre famille de mathématiciens suisses, séjourna à Saint-Petersbourg et y étudia le jeu suivant. Il s'agit en fait de lancer une pièce de monnaie jusqu'à l'obtention d'une première FACE. *Si le joueur lance une* FACE *au premier jet, il reçoit 1$. Sinon, pour le premier* PILE *obtenu, le joueur reçoit 2$ et la récompense est doublée pour chaque* PILE *subséquent. Si l'on dénote par n le nombre de fois que le joueur a ainsi obtenu* PILE, *le gain d'une joute est*

$$R(n) = 2^n \tag{3.10}$$

[†]Il peut être choquant de constater que l'on compare ici deux mesures qui ne sont pas dans les mêmes unités. Quoi qu'il en soit, cette fonction d'utilité est d'un usage très répandue (BODIE, KANE et MARCUS 2003).

[‡]Le facteur $\frac{1}{2}$ ne figure que pour enjoliver la dérivée de U.

et évidemment

$$\begin{aligned}\Pr(n) :&= \Pr(\textit{obtenir une première face } \textsc{face} \textit{ après avoir lancer } n \textit{ fois } \textsc{pile}) \\ &= \Pr(\textit{lancer } n \textit{ fois } \textsc{pile} \textit{ consécutivement}) \times \Pr(\textit{lancer } \textsc{face}) \\ &= \frac{1}{2^n} \times \frac{1}{2} = \frac{1}{2^{n+1}} \end{aligned} \tag{3.11}$$

Ainsi, l'espérance de gain est donc de

$$E[R] = \sum_{n=0}^{\infty} \Pr(n) R(n) \tag{3.12}$$

$$= \sum_{n=0}^{\infty} \frac{1}{2^{n+1}} 2^n \tag{3.13}$$

$$= \infty \tag{3.14}$$

Maintenant, quoique l'espérance de gain est infinie, si l'on demandait à un participant combien il prêt à payer pour jouer à ce jeu, il ne serait vraisemblablement prêt à payer qu'une somme finie, probablement même assez peu élevée. Ce jeu est appelé le paradoxe de Saint-Petersbourg.

Pour Bernoulli, la solution de ce paradoxe résidait dans une décroissance de la valeur perçue des gains supplémentaires, un concept que les économistes modernes appellent **décroissance de l'utilité marginale**. *Cette notion a été formalisée dans le cadre de la théorie de l'investissement en 1944 par John von Neumann et Oskar Morgenstern (*Morgenstern *et* von Neumann *1944).*

◊

L'utilité peut donc être vue comme une valeur subjective attribuée par un investisseur donné à un investissement. Notons que, peu importe A, un investissement dans le bien sans risque a une utilité U = E$[R_f]$ = r_f. Ainsi, un investisseur averse au risque rejetera d'emblée tout investissement dont la prime de risque est négative ou nulle.

Exemple 3.3 Investir ou non (suite)

Soucieux de son pouvoir d'achat sur le marché de l'immobilier à son retour,

monsieur X évalue son aversion au risque par $\mathrm{A} = 8$, *une valeur relativement élevée suivant la littérature économique. Ainsi, l'utilité de l'investissement considéré est*

$$\begin{aligned} \mathrm{U} &= \mathrm{E}[R] - \frac{1}{2}\mathrm{A}\sigma^2 \\ &= 0.13 - \frac{8}{2}0.0216 \\ &= 0.0436 \end{aligned}$$

L'investissement est donc moins intéressant qu'un investissement dans le bien sans risque qui a une utilité de $\mathrm{U} = r_f = 5\%$. *Cependant, si les 200 000$ provenaient d'un héritage plutôt que de la vente de sa maison, Monsieur X aurait probablement fixé son aversion au risque à* $\mathrm{A} = 4$, *une valeur modérée. Dans ce cas*

$$\begin{aligned} \mathrm{U} &= 0.13 - \frac{4}{2}0.0216 \\ &= 0.0868 \end{aligned}$$

auquel cas l'investissement dans les actions de ABC aurait été plus intéressant.

◇

Bref, l'aversion au risque n'est pas seulement propre à chaque investisseur, mais peut varier selon le contexte et le temps chez un même investisseur.

3.2 Sélection du portefeuille

Le but recherché par un gestionnaire de portefeuille est d'établir un portefeuille qui présente le meilleur compromis entre le rendement espéré et le risque encouru. La problématique attaquée dans cette section est le choix des biens à

introduire dans le portefeuille ainsi que le choix de leur pondération respective dans ledit portefeuille.

3.2.1 Ligne d'allocation du capital

D'entrée de jeu, spécifions que, dans la majorité des modèles, le bien sans risque tient une place à part.

Définition 3.4 *Un* **portefeuille risqué** P *est constitué uniquement d'actifs risqués, le bien sans risque n'entrant donc pas dans sa composition.*

Tout au long de ce mémoire, l'usage de P *référera à un portefeuille risqué,* R_P *et* r_P *désignant donc le rendement d'un tel portefeuille. Tout comme pour l'actif sans risque, l'ajout d'un indice de temps* t *(*R_{Pt} *ou* r_{Pt}*) souligne que l'on considére ledit rendement sur la période* t*.*

En effet, plusieurs modèle établisse d'abord un portefeuille risqué et répartisse ensuite leur investissement entre ce portefeuille et le bien sans risque afin d'obtenir un portefeuille présentant un risque cible.

Proposition 3.2.1 *(*Bodie, Kane *et* Marcus *2003) Soient* r_f *le taux sans risque sur une période donnée et* P *un portefeuille risqué auquel sont associés le rendement espéré* $\mathrm{E}[R_P]$ *et un écart-type* σ_P *sur la même période. Soit* C_y *un portefeuille constitué d'une proportion* y *du portefeuille risqué* P *et d'une proportion* $(1-y)$ *du bien sans risque.*

Si un investisseur désire obtenir un portefeuille C_y *de variance cible* σ_{C_y}*, alors la proportion à investir dans le portefeuille risqué est* $y = \sigma_{C_y}/\sigma_P$*.*

Preuve

$$\mathrm{Var}(R_{C_y}) = \mathrm{Var}((1-y)r_f + yR_P) \tag{3.15}$$

$$= y^2 \,\mathrm{Var}(R_P) \tag{3.16}$$

donc

$$y = \frac{\sigma_{C_y}}{\sigma_P} \tag{3.17}$$

■

Définition 3.5 CAL (*Capital Allocation Line*)
La **ligne d'allocation du capital** (Figure (3.1)) *est la ligne d'équation*

$$\mathrm{E}[R_{C_y}] = r_f + y(\mathrm{E}[R_P] - r_f) \tag{3.18}$$

$$= r_f + \left(\frac{\mathrm{E}[R_P] - r_f}{\sigma_P}\right)\sigma_{C_y} \tag{3.19}$$

qui associe à un risque cible σ_{C_y} le rendement attendu $\mathrm{E}[R_{C_y}]$ du portefeuille C_y.

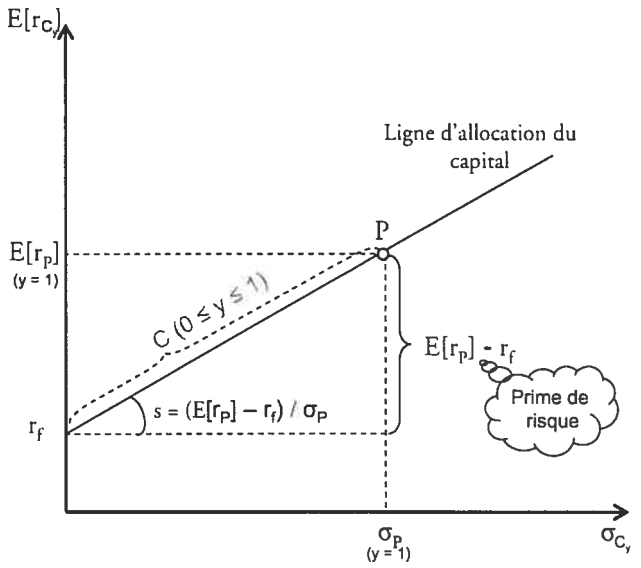

Figure 3.1 – *Ligne d'allocation du capital : cette droite permet, pour l'écart type cible σ_{C_y} d'un portefeuille complet C_y, de déduire le rendement espéré $\mathrm{E}[R_{C_y}]$ dudit portefeuille.*

Notons aussi que la pente $s = \frac{E[R_P]-r_f}{\sigma_P}$ est souvent appelée le **ratio de rendement associé au risque**, car elle représente le rendement supplémentaire obtenu par unité d'écart-type supplémentaire.

Définition 3.6 CML (*Capital Market Line*)
Soit M le **portefeuille de marché** — *un portefeuille risqué —, constitué d'un grand indice boursier donné et du bien sans risque. On appelle* **ligne de marché du capital** *la ligne d'allocation du capital associée à M.*

On dit passive une stratégie d'investissement qui consiste à établir son portefeuille selon la ligne de marché du capital, contrairement aux stratégies dites actives qui établissent leur portefeuille risqué suivant un modèle ou un autre. L'approche passive peut sembler naïve à prime abord, mais elle peut être viable.

D'abord, la stratégie passive est très peu coûteuse en frais de gestion : il suffit de payer les frais de transaction sur l'achat des bons du trésor et d'accepter qu'une faible partie de la somme investie dans le fond répliquant l'index aille à la gestion du fonds. Les stratégies actives, quant à elles, exigent soit du temps et des connaissances techniques poussés, soit le payement de frais de gestion autrement plus élevés.

Ensuite, si l'on accepte la théorie économique, les prix des biens en présence dans l'indice évoluent vers un équilibre où tout prix reflète justement la valeur et l'information disponible sur le bien. Cette évolution vers l'équilibre est entre autres dû à l'intervention des gestionnaires actifs qui, comme nous le verrons sous peu, tentent de profiter des inefficacités du marché. Ainsi, adopter une stratégie passive revient à faire confiance au travail acharné des requins de la finance et à profiter de l'évolution du marché qui en découle. Pour un petit investisseur aux connaissances économiques sommaires, une stratégie passive peut, en somme, être plus efficace que certaines stratégies actives.

De surcroît, les grands indices boursiers peuvent contenir jusqu'à plusieurs centaines des biens les plus fiables du marché. Ainsi, le portefeuille de marché jouie d'une excellente **diversification**.

3.2.2 Risque et diversification

Supposons qu'un investisseur désire se constituer un portefeuille risqué d'actions. Évidemment, s'il ne considère que les actions d'une compagnie, il s'expose à un risque qui est spécifique à cette compagnie. Si la compagnie vient à éprouver des difficultés majeures, son investissement en sera grandement affecté. Il est donc relativement intuitif de diversifier son portefeuille : plus notre investisseur détiendra d'actions de compagnies différentes et oeuvrant dans différents secteurs de l'économie, moins les **risques spécifiques** à chaque compagnie feront sentir leur influence respective dans le risque globalement encouru sur l'investissement.

Un portefeuille risqué de deux biens

Commençons avec un exemple simple ou un investisseur ne peut considérer que les deux biens 1 et 2 pour constituer son portefeuille. Si l'on dénote par w_1 et $w_2 = 1 - w_1$ les proportions respectives de 1 et 2 dans le portefeuille, alors le rendement du portefeuille P sera

$$R_P = w_1 R_1 + w_2 R_2 \tag{3.20}$$

dont l'espérance est

$$E[R_P] = w_1 E[R_1] + w_2 E[R_2] \tag{3.21}$$

et la variance

$$\sigma_P^2 = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 Cov(R_1, R_2) \tag{3.22}$$

$$= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 \sigma_1 \sigma_2 \rho_{12} \tag{3.23}$$

où ρ_{12} est la corrélation entre les rendements des biens 1 et 2.

Il est intéressant de remarquer que l'écart-type du portefeuille n'est pas, contrairement à l'espérance, une moyenne pondérée des écarts-type des rendements des deux biens en présence. Ainsi, l'équation 3.23 nous dit que l'écart-type du portefeuille sera moindre si la corrélation entre les rendements est négative. En fait, l'écart-type du portefeuille sera inférieure à la moyenne pondérée des écart-types de R_1 et R_2 à moins que les rendements de 1 et 2 ne soient parfaitement et positivement corrélés, c'est-à-dire que $\rho_{12} = 1$. Pour s'en convaincre, il suffit de constater que, pour $\rho_{12} = 1$, l'équation 3.23 devient

$$\sigma_P = w_1\sigma_1 + w_2\sigma_2 \tag{3.24}$$

la moyenne pondérée des écart-types. Comme σ_P^2, et par conséquent σ_P, est strictement croissante en ρ_{12}, si la corrélation est inférieure à 1, l'écart-type du portefeuille sera inférieure à la moyenne pondérée des écart-types de R_1 et R_2.

Étant donné que le rendement espéré d'un portefeuille est la moyenne pondérée des rendements espérés de chaque bien, mais que l'écart-type est inférieur à la moyenne pondérée des écarts-types, *un portefeuille établi avec des biens qui ne sont pas parfaitement et positivement corrélés offrira un meilleur compromis entre la moyenne et l'écart-type que chacun des biens pris individuellement.* Cette dernière observation constitue le fondement du principe de diversification.

Exemple 3.2.1 tiré de (BODIE, KANE et MARCUS 2003)
Supposons que l'on aie accès aux données suivantes

	Bien 1 (%)	**Bien 2 (%)**
$E[R]$	8	13
σ	12	20
ρ_{12}	0.3	

Alors, le rendement espéré et la variance du rendement du portefeuille seront

$$E[R_P] = 8w_1 + 13w_2 \tag{3.25}$$

$$\text{Var}(R_P) = 12^2 w_1^2 + 20^2 w_2^2 + 2 \times (12 \times 20 \times 0.3) w_1 w_2 \tag{3.26}$$

En résolvant le problème quadratique

$$\min_{w_1, w_2} \text{Var}(R_P) \tag{3.27}$$

$$w_1 + w_2 = 1 \tag{3.28}$$

on obtient

$$w_1^* = \frac{\text{Var}(R_2) - Cov(R_1, R_2)}{\text{Var}(R_1) + \text{Var}(R_2) - 2Cov(R_1, R_2)} \tag{3.29}$$

$$= 0.82 \tag{3.30}$$

$$w_2^* = 0.18 \tag{3.31}$$

Remarquons qu'en substituant ces poids dans l'éq. (3.26), le portefeuille P^* a un écart-type de 11.45%. L'écart-type du portefeuille ainsi obtenu est donc plus faible que celle des deux biens constituants le portefeuille.

L'exemple précédent permet de mettre en évidence l'effet de la diversification. Peu importe le nombre de biens considérés, le rendement espéré sera toujours la moyenne pondérée des rendements espérés de chaque bien. Or, à moins de n'avoir que des biens parfaitement positivement corrélés, la variance, quant à elle, sera inférieure à la moyenne pondérée des variances. Ainsi, un portefeuille *bien* diversifié présentera un meilleur compromis entre le rendement espéré et le risque encouru qu'un portefeuille constitué de trop peu de biens.

Quoi qu'il en soit, même un portefeuille extrêmement bien diversifié peut difficilement éliminer tout type de risque. En effet, certains facteurs macroéconomiques, par exemple, ont une influence non négligeable sur tout le spectre des investissements possibles. Il demeure donc un **risque de marché** ou **risque systématique** qui ne peut être diversifié. Par opposition, on appelle **risque**

non systématique le risque spécifique à chaque compagnie considérée par le portefeuille.

3.2.3 Modèles de sélection de portefeuille

Dans le dernier demi-siècle, nombre d'économistes se sont penchés sur le problème de la sélection de portefeuille. Il en résulta de nombreux modèles qui ont su, chacun à sa façon, faire avancer la théorie économique et ses applications. Que ce soit Harry M. Markowitz qui établit un des premiers modèles de la théorie du portefeuille moderne, modèle auquel il laissa son nom (MARKOWITZ 1952; MARKOWITZ 1991), ou Jack Treynor qui jeta les bases[§] (TREYNOR 1961) du CAPM (*Capital Asset Pricing Model*) dont l'influence théorique est encore aujourd'hui indéniablement présente, ou encore Stephen Ross qui donna naissance aux modèles reposant sur l'arbitrage (ROSS 1976b; ROSS 1976a), les grands noms de l'économie moderne se sont succédés dans l'étude de ce problème.

Sans entrer dans le détail, les outils théoriques nécessaires n'ayant pas été présentés, disons simplement que tous ces modèles tentent d'obtenir, d'une façon ou d'une autre, un portefeuille qui atteigne le meilleur compromis moyenne-variance. Un des enjeux majeurs de ces modèles de gestion active, est de tenter de profiter des inefficacités du marché.

[§]Le modèle a été complété par William Sharpe (SHARPE 1964), John Lintner (LINTNER 1965) et Jan Mossin (MOSSIN 1966) puis a été adapté à maintes reprises, entre autre par Fisher Black (BLACK 1964) et Eugene F. Fama (FAMA 1970).

Markowitz

Soit $V^{min}(c)$ la solution du problème suivant

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathrm{Var}(R_P) \\ & E[R_P] = c \\ & \sum_i w_i = 1 \end{aligned}$$

où $\mathbf{w}$ est le vecteur des poids de chaque bien dans un portefeuille risqué P.

En résolvant ce problème pour tout c raisonnable, on obtient

$$V^{min} = \{V^{min}(c) \,|\, c \in \mathcal{C} \subset \mathbb{R}\}, \tag{3.32}$$

la **frontière de variance minimum**.

Soit $\tilde{c} = \arg\min_c V^{min}(c)$, les économistes appellent **frontière efficiente** la partie de la frontière de variance minimum qui se trouve au-dessus de $\tilde{c}$, car pour tout portefeuille sur la partie inférieure de la courbe, il existe un portefeuille de variance égale, mais de rendement supérieur, qui se trouve sur la frontière efficiente (Bodie, Kane et Marcus 2003). Les portefeuilles de la frontière efficiente **dominent** donc ceux de la partie inférieure de la frontière de variance minimum (voir Figure (3.2)).

Une fois la frontière efficiente établie, il faut choisir un des portefeuilles risqués dont la variance repose sur ladite frontière. Le modèle de Markowitz (Markowitz 1952) suggère de choisir le portefeuille risqué P dont la variance correspond au point de tangence entre une ligne d'allocation du capital (Définition (3.5)) et la frontière efficiente. En effet, tout autre portefeuille risqué sur la frontière efficiente est strictement dominé par un portefeuille complet sur la CAL composé de P et de r_f (Figure (3.2) ; image de droite).

L'investisseur pourra ensuite choisir un portefeuille complet suivant l'exposé de la section 3.2.1.

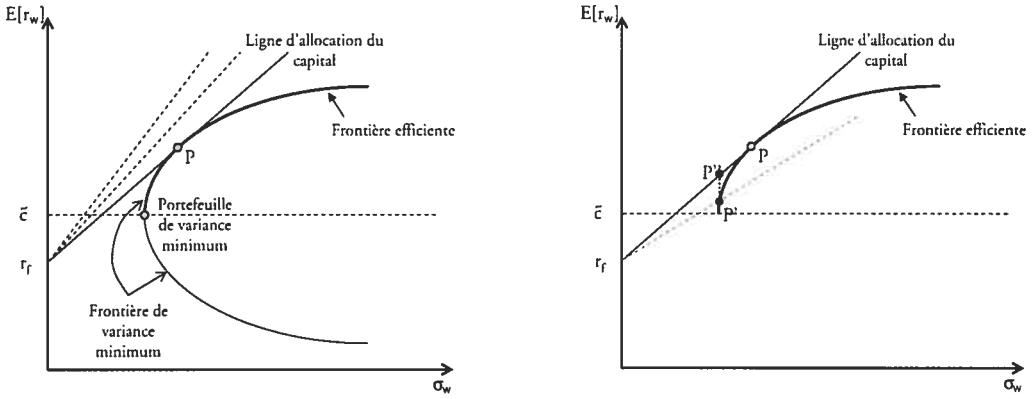

Figure 3.2 – *Frontière efficiente et ligne d'allocation du capital.*
L'abscisse résume un portefeuille $\mathbf{w}$ *par sa variance* $\sigma_{\mathbf{w}}$. *La parabole constitue la frontière de variance minimum* V^{min}, *laquelle n'est constituée que de portefeuilles risqués. Or, tout portefeuille sur la partie inférieure de la parabole est strictement dominé par un portefeuille sur la partie supérieure. Conséquemment, cette partie supérieure est dite la frontière efficiente.*
*Une fois la frontière efficiente établie, il faut choisir un des portefeuilles risqués dont la variance repose sur ladite frontière. Le modèle de Markowitz (*MARKOWITZ *1952) suggère de choisir le portefeuille* P *dont la variance correspond au point de tangence entre une ligne d'allocation du capital (Définition (3.5)) et la frontière efficiente.*
En effet, si l'on choisissait (figure de droite) un autre portefeuille P', *il serait toujours possible de trouver, sur la CAL reliant* r_f *et* P, *un portefeuille complet* P'' *qui dominerait* P' *au sens où, présentant le même risque que* P', P'' *présente aussi un plus grand rendement espéré.*

Le CAPM

Le CAPM (*Capital Asset Pricing Model*; (SHARPE 1964; LINTNER 1965; MOSSIN 1966)) tente de prédire le rendement que devrait effectivement présenter un bien compte tenu du risque qui lui est associé (tiré des donnés historiques) et de la valeur de ce risque dans le contexte actuel du marché. Un investisseur peut ensuite acheter ou vendre un bien suivant qu'il présente un rendement plus ou moins élevé que celui prédit par le CAPM.

Comme nous l'avons mentionné au début du chapitre (§ 3.2.1), une stratégie passive consistant à investir dans un portefeuille de marché est efficiente au sens qu'elle sous-tend un rendement espéré maximal compte tenu du risque encouru et ce en vertu d'une diversification avisée. Soit $(\mathrm{E}[R_M] - r_f)$ la prime de risque de la stratégie passive. Alors, intuitivement, tout investissement dans un bien risqué devrait avoir une prime de risque proportionnelle à celle de la stratégie passive. En clair, si β_k décrit la relation entre le risque sur un bien k et le risque sur le portefeuille de marché, alors

$$\mathrm{E}[R_k] - r_f = \beta_k(\mathrm{E}[R_M] - r_f) \tag{3.33}$$

$$\Leftrightarrow \quad \mathrm{E}[R_k] = r_f + \beta_k(\mathrm{E}[R_M] - r_f) \tag{3.34}$$

Pour caractériser la relation entre les risques en présence, les économistes utilisent $\beta_k = \frac{\mathrm{Cov}(R_k, R_M)}{\sigma_M^2}$, la contribution du risque sur le bien k au risque global du marché.

Du point de vue théorique, le CAPM repose sur plusieurs hypothèses plus ou moins réalistes qui ne survivent pas nécessairement bien à toutes les études (Fama et French 1992). Quoi qu'il en soit, il demeure une pièce centrale de l'économie financière moderne, car il permet de développer une intuition sur plusieurs problèmes pratiques et est suffisamment efficace (Mayers 1972; Amihud, Bent et Mendelson 1992) pour permettre certaines prédictions qui vont bien plus loin que la simple gestion de portefeuille.

L'arbitrage

On appelle **arbitrage** l'élaboration d'une stratégie d'achat et de vente de biens corrélés de telle façon que le risque encouru par la stratégie est nul bien que le rendement promis soit positif. Lorsque l'élaboration d'une telle stratégie est possible, il est évident que le marché démontre une inefficacité dans l'établissement du prix d'au moins un des biens en présence.

Exemple 3.4

Dans un monde simplifié, où il n'y a que trois scénarios économiques possibles, récession, normal et boom, supposons que les donnés historiques sur trois actions (A, B et C) nous permettent de croire aux comportement suivants

		Rendement sous le scénario (%)		
Action	Coût	Récession	Normal	Boom
A	*10*	*-15*	*20*	*30*
B	*15*	*25*	*10*	*-10*
C	*50*	*12*	*15*	*12*

Maintenant, supposons que l'on vende deux actions A, deux actions B et que l'on achète simultanément une action C

		Gains suivant le scénario ($)		
Action	Investissement ($)	Récession	Normal	Boom
A	*-20*	*3*	*-4*	*-6*
B	*-30*	*-7.5*	*-3*	*3*
C	*+50*	*6*	*7.5*	*6*
	0$	*1.5$*	*0.5$*	*3$*

L'espérance de gain est évidemment positive, bien que l'investissement initial est nul.

◇

Certains investisseurs, communément appelés **arbitrageurs** sont sans cesse à l'affût de telles inefficacités sur le marché. Un des principes fondamentaux de la théorie des marchés financiers veut que les prix évoluent de telles façon à ce que possibilité d'arbitrage ne soit que passagère, les prix évoluant vers un équilibre où aucun arbitrage n'est possible. Intuitivement, il apparaît évident que si une possibilité d'arbitrage est détectée et utilisée par un nombre significatif d'investisseurs, la stratégie, que l'on peut voir comme un portefeuille dont

la valeur est sous-évaluée, accusera une demande croissante avec le nombre d'arbitrageurs qui l'adopterons. Sous l'influence de cette croissance de la demande, le prix de la stratégie augmentera de telle façon que la sous-évaluation disparaîtra. D'une certaine façon les arbitrageurs assurent donc un maintien de l'équilibre du marché, travail qui selon plusieurs justifient le salaire touché par le biais des profits de leur pratique.

Partie III

L'APPRENTISSAGE AUTOMATIQUE

L'apprentissage automatique[a] est un domaine aux influences diverses. Des statistiques à la recherche opérationnelle, en passant par la psychologie cognitive, moult sciences s'allient dans cette tentative d'implanter chez la machine une faculté qui fut centrale au développement de l'humanité : la capacité d'apprendre et de généraliser. En effet, au cœur même de l'évolution de l'être humain se trouve sa capacité à s'adapter à son environnement, d'apprendre de ses erreurs, de savoir reconnaître dans de nouvelles situations quelques similitudes avec une expérience précédente et de savoir en tirer parti.

Au sens large, toute méthode qui incorpore l'information donnée de manière à en déduire un aspect non trivial permettant de prendre des décisions futures est un algorithme d'apprentissage. Ainsi, le calcul d'une simple moyenne mobile permet d'**apprendre** une partie du comportement des données à travers le temps. Aussi vaste que puisse être ce domaine, il peut être segmenté en grandes catégories desquelles nous citerons l'**apprentissage supervisé**, l'**apprentissage non supervisé** et l'**apprentissage semi-supervisé**. Le chapitre 4 se veut une introduction à la première catégorie et le suivant à la seconde. Pour ce qui est de l'apprentissage semi-supervisé, comme ce mémoire n'en fait pas un usage direct, nous nous contenterons de l'introduire brièvement à la fin du chapitre 5.

[a] Anglais : *Machine Learning*

CHAPITRE 4

Apprentissage supervisé

Comme le nom le laisse pressentir, dans le cadre d'un apprentissage supervisé, le modèle dispose en quelque sorte d'un professeur. En fait, à chaque exemple fourni pour l'apprentissage sera associée une étiquette qui sera soit la réponse exacte que l'on s'attend que le modèle nous procure, soit une information suffisamment forte pour guider le modèle dans son apprentissage.

4.1 Représentation des données

On considérera un **ensemble d'entraînement** $\mathcal{D}$ comme un ensemble

$$\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \,|\, i = 1, \ldots, m\} \tag{4.1}$$

où $\mathcal{X}$ est l'ensemble non vide des valeurs possibles pour les **observations**, $\mathcal{Y}$ les **étiquettes** possibles et m la taille de l'ensemble d'entraînement. Il est à noter que $\mathcal{X}$ et $\mathcal{Y}$ sont quelconques : les observations et les étiquettes peuvent être scalaires, vectorielles ou même non numériques.

4.2 Tâches d'apprentissage supervisé

Nous introduirons les deux types les plus communs d'apprentissage supervisé, soient la **classification** et la **régression**. Dans les deux cas, il faut avoir conscience que les étiquettes sont possiblement bruitées. En effet, les bases de données considérées ne sont pas immunisées contre les erreurs humaines ou contre de simples erreurs de mesures dues aux limites technologiques. Quoiqu'il en soit, nous supposons qu'il existe une fonction $f^* : \mathcal{X} \to \mathcal{Y}$ telle que

$$y = f^*(x) + \epsilon, \tag{4.2}$$

et qui permette d'obtenir la meilleure classification/régression possible sur les paires (x, y) affectées de bruits ϵ aléatoires d'espérance nulle.

Le défi est donc d'approximer, étant donné un ensemble d'entrainement $\mathcal{D}$, la fonction f^* à l'aide d'une fonction

$$f : \mathcal{X} \to \mathcal{Y} \tag{4.3}$$

$$x \mapsto f(x) \tag{4.4}$$

telle que f soit le plus près* possible de f^*.

4.2.1 Classification

Dans le cadre de la classification, $\mathcal{Y}$ est un ensemble fini ou dénombrable de classes, c'est-à-dire

$$\mathcal{Y} = \{\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_i, \ldots\} \tag{4.5}$$

*Déjà, on pressent une notion d'erreur ou de distance dans l'espace des fonctions qui peut être définie différemment selon la tâche à résoudre.

Le but de l'apprentissage est alors de prévoir la classe Y_i à laquelle appartient une observation $x \in \mathcal{X}$.

Exemple 4.1 Un cas simple — 2 Classes
Soit un ensemble d'extraits de solutions impures d'acide acétique et d'acide citrique à 20% de masse

$$\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \,|\, i = 1, \ldots, 20\} \tag{4.6}$$

tel que

$$x_i = (\textit{masse volumique de l'extrait } i \textit{ en } g/cm^3, \tag{4.7}$$
$$\textit{viscosité de l'extrait } i \textit{ en } mPa{\times}s) \tag{4.8}$$
$$\mathcal{X} = \mathbb{R}_+ \times \mathbb{R}_+ \tag{4.9}$$

et

$$y_i = \textit{type d'acide dans la solution} \tag{4.10}$$
$$\mathcal{Y} = \{\text{ACÉTIQUE}, \text{CITRIQUE}\} \tag{4.11}$$

Étant donné les exemples fournis par $\mathcal{D}$, si l'on désire classer de façon automatique un nouvel échantillon, on peut se contenter de tracer une droite qui sépare les deux groupes et d'affecter à la nouvelle instance l'étiquette correspondant à sa position relative à ladite droite (figure 4.1).

Ainsi, la fonction de classification pourrait être (Fig. (4.1)), pour $x = (\rho, \eta)$

$$f(x) = \begin{cases} \text{ACÉTIQUE} & : \quad si\ \rho >= -0.6111\eta + 2.2303 \\ \text{CITRIQUE} & : \quad si\ \rho < -0.6111\eta + 2.2303 \end{cases} \tag{4.12}$$

$\diamondsuit$

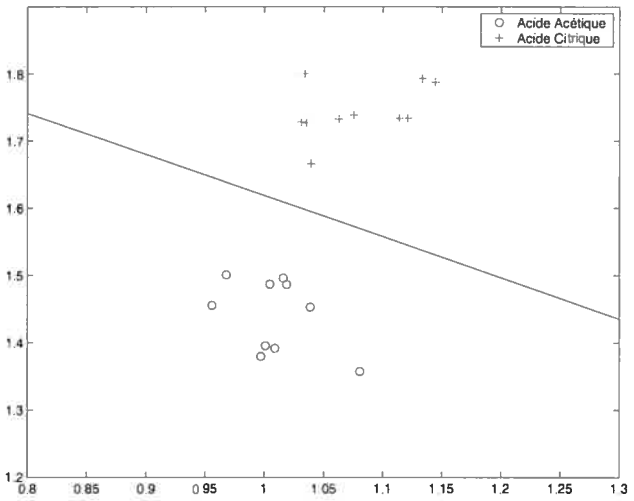

Figure 4.1 – *Classification : Sur les exemples provenant de $\mathcal{D}$, la droite d'équation $\rho = -0.6111\eta + 2.2303$ sépare parfaitement les deux types d'échantillons.*

4.2.2 Régression

Dans un contexte de régression, $\mathcal{Y}$ est habituellement un ensemble continu de valeurs.

Exemple 4.2 Régression Linéaire (MOORE et MCCABE **1999**)
L'indice lipidique, représentant la proportion de graisses contenue dans les tissus d'un individu, est d'une utilité médicale indéniable. Il permet entre autre de pressentir des problèmes cardiaques ou de doser la quantité d'analgésique nécessaire à l'anesthésie d'un patient. Pour obtenir l'indice de façon exacte, il faut en fait évaluer la densité corporelle du patient. Or, pour obtenir cette densité corporelle, la procédure première nécessite une pesée sous l'eau. Il va sans dire qu'en situation d'urgence ou même pour avoir une idée rapide de l'indice lipidique d'un patient, cette procédure est loin d'être pratique.

Pour remédier à cette fâcheuse situation, les chercheurs ont suggéré une approche qui consiste à pincer la peau en quatre endroits stratégiques et à noter l'épaisseur de la peau. On prend ensuite le logarithme de la somme de ces

quatre mesures. Le résultat ainsi obtenu est dit l'indice LSKIN[†]. *Pour prédire la densité corporelle d'un patient étant donné l'indice* LSKIN *les chercheurs ont constitué*[‡] *l'ensemble d'entraînement suivant*

$$\begin{aligned}\mathcal{D} &= \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, 92\} \\ &= \{(1.27, 1.093), (1.56, 1.063), \ldots, (1.52, 1.056)\}\end{aligned}$$

tel que

$x_i =$ *indice* LSKIN *du patient* i

$\mathcal{X} = \mathbb{R}^+$

et

$y_i =$ *densité corporelle* i

$\mathcal{Y} = \mathbb{R}^+$

Il suffit de jeter un coup d'œil au graphe (figure (4.2)) pour suspecter l'efficacité d'un prédicteur linéaire.

La droite d'équation $y = -0.0631x + 1.163$ *est celle qui approxime le mieux*[§] *la relation entre l'indice* LSKIN *et la densité corporelle. Ainsi, si un patient présente un indice* LSKIN *de 1.27, on lui attribuera une densité approximative* $f(x) = 1.0999$, *ce qui, pour le patient 1, aurait représenté une erreur d'à peine 0.6%.*

◇

4.3 Minimisation de l'erreur empirique

Dans les exemples précédents, une fois les données posées, nous donnons d'emblé le prédicteur qui fait le travail. Comme ces exemples étaient fort

[†] *De l'anglais : Log Skinfold Thickness*

[‡] *La densité corporelle a été mesurée suivant la procédure d'immersion.*

[§] *Cette locution laisse pressentir la notion de critère d'optimisation, notion traitée à la section 4.3.*

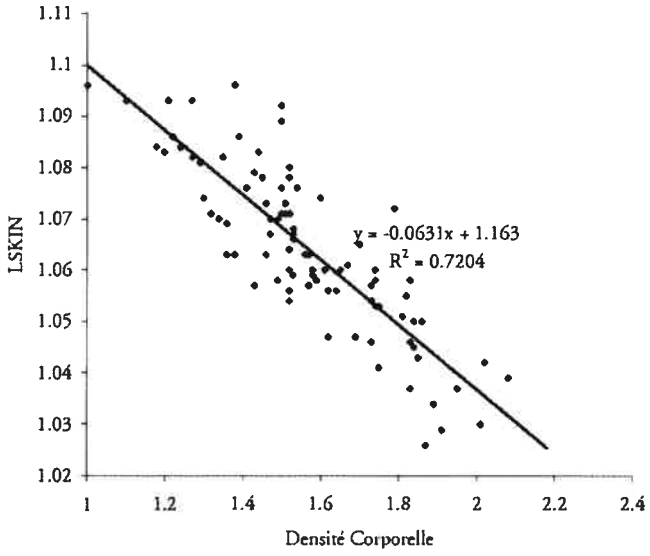

Figure 4.2 – *Régression Linéaire : Prédire la densité corporelle à l'aide de la mesure* LSKIN.

simples, leur solution était assez intuitive. Or, dès que nous considérerons des problèmes de plus haute dimension, l'intuition graphique ne sera plus à notre portée.

Comme nous l'avons déjà dit, l'apprentissage consiste en somme à approcher la fonction f^* inconnue qui sous-tend les données par une fonction f estimée à partir des données d'entraînement. Les questions soulevées dans les prochaines sections sont à savoir comment trouver cette fonction et comment évaluer sa qualité.

La première question qui s'impose est en fait : "Comment apprend-on ?".

Définition 4.1 *Étant donnée* $f \in \mathcal{F}_{\mathcal{X}\to\mathcal{Y}}$, *où* $\mathcal{F}_{\mathcal{X}\to\mathcal{Y}}$ *est l'ensemble qui contient toute fonction de* $\mathcal{X}$ *dans* $\mathcal{Y}$[¶], *on appelle* **erreur d'entraînement** *ou* **erreur**

[¶] *Évidemment,* $\mathcal{F}_{\mathcal{X}\to\mathcal{Y}}$ *est extrêmement vaste. Nous reviendrons sur l'obligation de se restreindre à un sous-ensemble de* $\mathcal{F}_{\mathcal{X}\to\mathcal{Y}}$ *(§ 4.4.1).*

empirique *la fonction*

$$(\mathcal{L}_{\mathcal{D}} f) : (\mathcal{X} \times \mathcal{Y}^2)^m \to \mathbb{R} \tag{4.13}$$

$$\{(x_i, y_i, f(x_i))\}_{(x_i,y_i)\in\mathcal{D}} \mapsto \frac{1}{m}\sum_{i=1}^{m} c(x_i, y_i, f(x_i)) \tag{4.14}$$

où $c : \mathcal{X} \times \mathcal{Y}^2 \to \mathbb{R}$ *est la* **fonction de coût** *qui évalue le coût encouru par une prédiction* $f(x)$ *sachant que l'étiquette réellement associé à* x *est* y*. On requerra habituellement que* $c(x, y, y) = 0, \forall x \in X, y \in Y$*.*

Ainsi, dans un cadre d'apprentissage supervisé, la réponse à la question précédente passe plus souvent qu'autrement par la minimisation de l'erreur empirique[‖], c'est à dire

$$f \in \underset{\underline{f}\in\mathcal{F}_{\mathcal{X}\to\mathcal{Y}}}{\arg\min}(\mathcal{L}_{\mathcal{D}}\underline{f}) \tag{4.16}$$

Quoi qu'il en soit, il est à noter que pour une fonction c donnée et $m < \infty$, une erreur empirique faible ne nous assure pas nécessairement de la qualité du prédicteur.

Exemple 4.3 Un cas simple — 2 Classes (suite)
Revisitons l'exemple de classification des solutions d'acides acétique et citrique. Soit c_1 *la fonction de coût telle que*

$$c_1(x, y, f(x)) = \begin{cases} 0 & : \quad si\ f(x) = y \\ 1 & : \quad si\ f(x) \neq y \end{cases} \tag{4.17}$$

[‖]Cette approche découle du principe inductif de minimisation structurelle du risque introduite dans (VAPNIK 1995). Le lecteur comprendra que le choix de minimiser *s'accorde* avec la terminologie *erreur empirique*. Si un problème requérait la maximisation des **contributions** c, il suffirait d'avoir recours à l'identité

$$\max_{\underline{f}} (\mathcal{L}_{\mathcal{D}}\underline{f}) = -\min_{\underline{f}} -(\mathcal{L}_{\mathcal{D}}\underline{f}). \tag{4.15}$$

Alors, il existe une infinité de prédicteurs qui annulent l'erreur empirique ; à celui déjà donné, nous en ajoutons ici un second en pointillés (figure (4.3)).

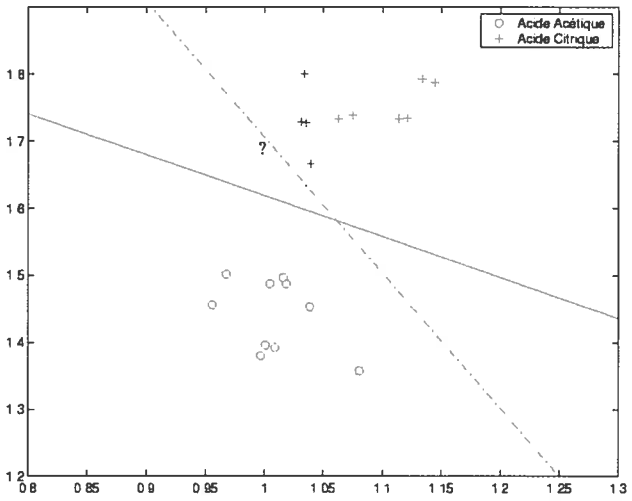

Figure 4.3 – *Les deux droites permettent d'annuler l'erreur empirique. Quoiqu'il en soit, l'une nous semble intuitivement meilleure que l'autre...*

Cependant, il suffit de tenter de classer le point identifié par le symbole '?' pour réaliser la piètre qualité du nouveau prédicteur. En effet, il est relativement évident que le nouvel échantillon est une solution d'acide citrique. Ainsi, contrairement au premier prédicteur, le second ne prédit pas correctement l'étiquette du nouvel exemple.

Par ailleurs, si nous avions utilisé une fonction de coût c_2 telle que

$$c_2(x, y, f(x)) = \begin{cases} -\,\mathrm{d}(x, \mathrm{proj}_D(x)) & : \quad si\ f(x) = y \\ \mathrm{d}(x, \mathrm{proj}_D(x)) & : \quad si\ f(x) \neq y \end{cases} \tag{4.18}$$

où proj_D est la projection (Définition (1.13)) sur la droite D d'équation $\rho = -0.6111\eta + 2.2303$, alors nous aurions eu pour seule solution le premier prédicteur.

$\Diamond$

4.4 Généraliser !

Non seulement l'exemple précédent démontre-t-il l'importance de bien choisir la fonction de coût, mais il introduit aussi l'enjeu principal de l'apprentissage automatique : la généralisation à de nouveaux cas. Il va sans dire, l'entraînement d'un prédicteur ne prend vraiment tout son sens que s'il permet d'utiliser ce dernier pour prédire, lorsque confronté à un nouvel exemple x, l'étiquette y qui lui est associée.

Or, si l'on nous demandait lequel des deux prédicteurs obtenus à l'exemple 4.3 privilégier, il semble que l'erreur empirique ne nous serait d'aucune utilité, les deux prédicteurs présentant une erreur empirique nulle. Il faut donc recourir à d'autres concepts.

Définition 4.2 *Supposons que les données soient générées par une loi de probabilité μ fixe mais inconnue, alors l'***erreur de généralisation** *d'un prédicteur f étant donnée une fonction de coût c est donnée par*

$$(\mathcal{L}f) = \int c(x, y, f(x))\ d\mu(x, y) \tag{4.19}$$

advenant bien sûr que c soit telle que l'intégrale converge.

La raison pour laquelle l'erreur empirique peut nous induire en erreur est simple. Il s'agit en fait d'un estimateur fortement bruité de l'erreur de généralisation. En effet, comme l'optimisation du prédicteur est dirigée suivant $\mathcal{D}$, il est possible que notre prédicteur surapprenne.

Pour en venir à une comparaison valable de différents modèles, la notion suivante est donc indispensable.

Définition 4.3 *Étant donné un ensemble*

$$\mathcal{T} = \{(\hat{x}_i, \hat{y}_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, m_{\mathcal{T}}\} \tag{4.20}$$

disjoint de $\mathcal{D}$ et f un prédicteur donné, on appelle **erreur de test** *sur $\mathcal{T}$*

$$(\mathcal{L}_{\mathcal{T}} f) : (\mathcal{X} \times \mathcal{Y}^2)^{m_{\mathcal{T}}} \to \mathbb{R} \tag{4.21}$$

$$\{(\hat{x}_i, \hat{y}_i, f(\hat{x}_i))\}_{(\hat{x}_i,\hat{y}_i)\in\mathcal{T}} \mapsto \frac{1}{m_{\mathcal{T}}} \sum_{i=1}^{m_{\mathcal{T}}} c(\hat{x}_i, \hat{y}_i, f(\hat{x}_i)). \tag{4.22}$$

4.4.1 Capacité et régularisation

Jusqu'à maintenant, nous avons passé sous silence un élément important de l'apprentissage : la **capacité** (Vapnik et Chervonenkis 1968; Vapnik et Chervonenkis 1971). Peu importe la fonction de coût choisie, avant de pouvoir trouver le prédicteur f qui minimise l'erreur empirique, il faut déterminer à quelle classe de fonctions $\mathcal{F}_\Theta$ limiter la recherche.

Lorsque l'on recherche une fonction dans un espace donné, on recherche en fait les paramètres $\theta \in \Theta$ qui caractérisent la fonction dans cet espace, c'est à dire θ tel que $f_\theta(x) = f(x;\theta)$. Pour une régression linéaire, par exemple, on cherchera $\theta = (a, b)$ tel que $y = ax + b$, tandis que pour une régression exponentielle on cherchera $\theta = (a, b)$ tel que $y = ae^{bx}$.

Ainsi, on peut réécrire l'équation 4.16 comme étant

$$f \in \arg\min_{\theta\in\Theta}(\mathcal{L}_{\mathcal{D}} f_\theta). \tag{4.23}$$

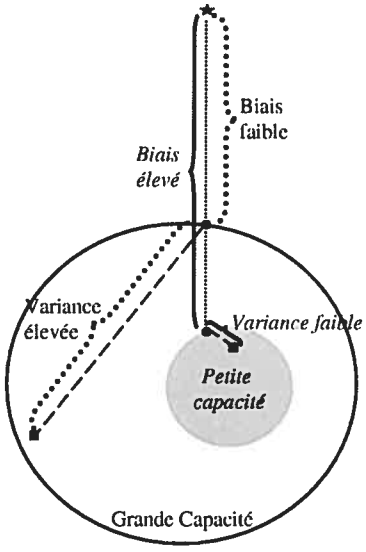

Figure 4.4 – *La fonction optimale f^* est représentée ici par une étoile. La projection de f^*, représentée par un cercle, est le meilleur estimateur que l'on puisse atteindre dans l'espace en question. Dans l'espace, dépendemment de l'ensemble d'entraînement, on apprendra f, ici en carrée.*

$$\textit{Biais}^2 : \quad \left\| \mathrm{E}\left[f \mid \mathcal{D}\right] - (\mathcal{L}f^*) \right\|^2$$

$$\textit{Variance} : \quad \mathrm{E}\left[\left\| f - \mathrm{E}\left[f \mid \mathcal{D}\right] \right\|^2 \right]$$

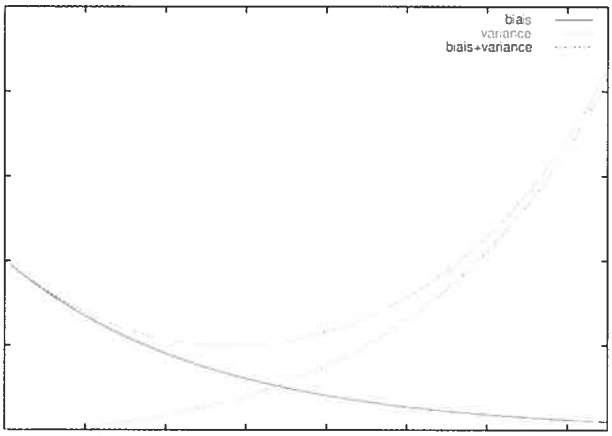

Figure 4.5 – *Capacité et surapprentissage : Plus la capacité d'un modèle augmente et plus son biais est faible, c'est-à-dire que l'espérance de la fonction empirique en fonction des données sera plus proche de la fonction réelle. Cependant, sa variance croîtra avec la capacité, l'espérance de l'écart (au carré) entre la fonction empirique obtenue et l'espérance de cette fonction empirique augmentant.*

La somme du biais et de la variance est une fonction convexe de la capacité. Le choix d'une capacité trop faible entraînera un **sous-apprentissage**, *un capacité trop grande, un* **surapprentissage**.

D'autre part, si nous considérons un espace de fonctions $\mathcal{F}_\Theta$ trop petit, il est fort probable que f^* ne se trouve pas dans $\mathcal{F}_\Theta$, entraînant ainsi un **biais** systématique. Par ailleurs, si pour remédier à cette situation, nous décidons d'attaquer un espace de plus grande taille, d'une plus grande *capacité**, l'optimisation risque d'être très sensible à l'ensemble d'entraînement $\mathcal{D}$, c-à-d. qu'il sera enclin au surapprentissage (voir Figure (4.5)). Ainsi, f (Éq. (4.23)) aura une grande **variance** (Voir figures 4.4 à 4.6).

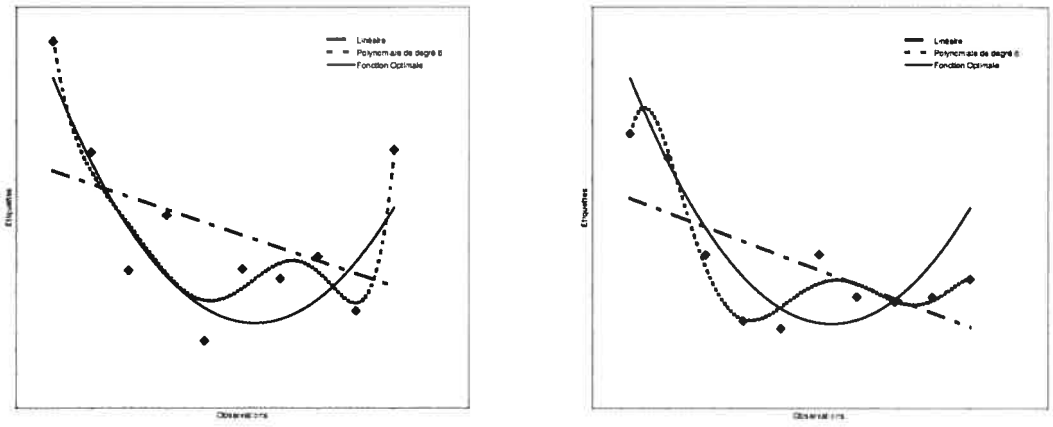

Figure 4.6 – *La même fonction optimale, mais deux ensembles de données bruitées différents. Dans les deux cas, la fonction optimale est approximée par des régressions polynomiales de degré 1 et 6. La capacité expressive des polynômes de degrés 1 est trivialement moins grande que celle des polynômes de degré 6 et c'est pourquoi l'***erreur d'approximation** *de la régression linéaire est plus grande. Or, l'***erreur d'estimation** *de cette dernière est beaucoup moins grande que celle de la régression d'ordre 6 qui est très sensible à l'ensemble d'entraînement.*

Pour remédier au dilemme de la capacité, on laissera souvent au modèle la possibilité d'avoir recours à une plus grande capacité moyennant une pénalisation *proportionnelle* à la capacité utilisé. La méthode de **régularisation**, proposée par A. N. Tikhonov en 1963, reprise plus tard dans (TIKHONOV et ARSENIN 1977), permet de formaliser cette idée.

*Notre approche du concept de capacité se veut plus intuitive que rigoureuse. Pour une étude irréprochable du concept, le lecteur peut se référer, entre autres, à (VAPNIK 1995; VAPNIK 1998; SCHÖLKOPF et SMOLA 2002)

Définition 4.4 *Étant donnée* $(\mathcal{L}_{\mathcal{D}} f)$, *on définit l'*__erreur régularisée__ *comme étant*

$$(\mathcal{L}_{\mathcal{D}}^{\Omega} f) = (\mathcal{L}_{\mathcal{D}} f) + \Omega(f) \tag{4.24}$$

où $\Omega : \mathcal{F}_{\Theta} \to \mathbb{R}^{+}$ *une* **fonction de pénalisation** *telle que les ensembles*

$$\mathcal{M}_{\nu} = \{f : \Omega(f) < \nu\}, \quad \nu \geq 0 \tag{4.25}$$

soient tous compacts[†].

Remarquons que le choix de Ω est très important puisque l'optimisation de $(\mathcal{L}_{\mathcal{D}}^{\Omega} f)$ y est très sensible. Son choix doit donc être fait avec soin. Entre autres méthodes pour discriminer entre les différentes fonctions possibles, plusieurs utilisent la validation croisée (voir (Stone 1974; Golub, Heath et Wahba 1979; Golub et von Matt 1997)).

[†] *La compacité des ensembles* $\mathcal{M}_{\nu}$ *est requise pour établir certains théorèmes de convergences (*Vapnik *1998).*

CHAPITRE 5

Apprentissage non supervisé

Dans le cadre de l'apprentissage non supervisé, l'algorithme ne dispose d'aucune information quant à la sortie qu'il devrait produire. Typiquement, ce cadre regroupe des techniques d'**estimation de densité**, de **partitionnement*** et de **réduction de dimensionnalité**. Les premières cherchent essentiellement à modéliser la distribution de laquelle ont été tirées les données. Pour sa part, le partitionnement est très semblable à la classification, au sens que les données proviennent de différents groupes. À la différence de son pendant supervisé, le partitionnement ne dispose ni de données bien classifiées pour l'entraînement, ni même du nombre de partitions desquelles sont issues les données. Finalement, les méthodes de réduction de la dimensionalité recherchent, dans une espace de haute dimension, une variété de plus faible dimension telle que la projection des points sur cette variété permette de conserver l'essentiel de l'information originale.

*En anglais : *clustering*

5.1 Représentation des données

Ici, nous considérons un **ensemble d'entraînement** $\mathcal{D}$ comme étant un ensemble

$$\mathcal{D} = \{x_i \in \mathcal{X} \mid i = 1, \ldots, m\} \tag{5.1}$$

où $\mathcal{X}$ est l'ensemble non vide des valeurs possibles pour les observations. Encore une fois, il est à noter que $\mathcal{X}$ est quelconque : les observations pouvant être scalaires, vectorielles ou même non numériques. Remarquons que cette représentation est tout à fait cohérente avec la représentation présentée à la section 4.1 qui n'excluait en rien que $\mathcal{Y}$ soit vide.

Évidemment, sous ces conditions, un apprentissage découlant directement de l'optimisation d'une fonction de coût n'est plus nécessairement intuitive. La grande majorité des algorithmes non supervisés sont plutôt inspirés par des intuitions géométriques propres à chaque problème. La prochaine section présente ainsi une technique de réduction de la dimensionnalité d'utilité dans ce mémoire dont l'inspiration, comme le lecteur saura le constater, est purement géométrique.

5.2 Analyse en composantes principales

L'analyse en composantes principales, ACP[†], est une technique classique d'extraction des caractéristiques. L'ACP n'est en fait qu'une transformation orthogonale du système de coordonnées dans lequel les données sont décrites. Soit un ensemble d'observations

$$\tilde{\mathcal{D}} = \{\tilde{x}_i \in \mathbb{R}^n \mid i = 1, \ldots, m\}$$

[†]Anglais : *Principal Component Analysis (PCA)*

de moyenne $\overline{x}$. L'ACP considérera les données centrées

$$\mathcal{D} = \{x_i = \tilde{x}_i - \overline{x} \in \mathbb{R}^n \mid i = 1, \ldots, m\}$$

où l'on remarque que $\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} (\tilde{x}_i - \overline{x}) = (\sum_{i=1}^{m} \tilde{x}_i) - m\overline{x} = 0$. Définissons maintenant la matrice de covariance sur les données centrées comme étant

$$\mathrm{Cov} = \mathrm{E}_x[xx^{\mathrm{T}}],$$

soit l'espérance de la matrice symétrique xx^{T} étant donnée la distribution de x. N'ayant évidemment pas accès à cette distribution, nous considérerons un estimateur de Cov sur $\mathcal{D}$

$$\mathrm{C} = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^{\mathrm{T}}. \tag{5.2}$$

Notons que C est symétrique et définie positive (Définition (1.16)) et qu'il existe donc un ensemble de m vecteurs propres normalisés v_k, tous non nuls, associés aux valeurs propres réelles $\lambda_k \geq 0$ tels que

$$\mathrm{C}v_k = \lambda_k v_k \qquad\qquad k = 1, \ldots, m \tag{5.3}$$

où nous considérons les valeurs propres en ordre décroissant, i.e. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. Enfin, nous noterons V_K la matrice $(K \times n)$ dont les lignes sont les vecteurs propres, dits **principaux**, associés au K plus grandes valeurs propres de C.

Pour tout K, il est possible de démontrer (Diamantaras et Kung 1996) que la projection

$$y_i = \mathrm{V}_K \, x_i \tag{5.4}$$

en est une qui minimise l'erreur quadratique de reconstruction d'une projection en dimension K, c'est-à-dire que

$$V_K \in \arg\min_W \mathrm{E}[\|x - W^{\mathrm{T}} W x\|] \tag{5.5}$$

$$WW^{\mathrm{T}} = I, \text{l'identité en dimension } K \tag{5.6}$$

$$\mathrm{span}\{W\} = \mathbb{R}^k \text{(Définition (1.2))}. \tag{5.7}$$

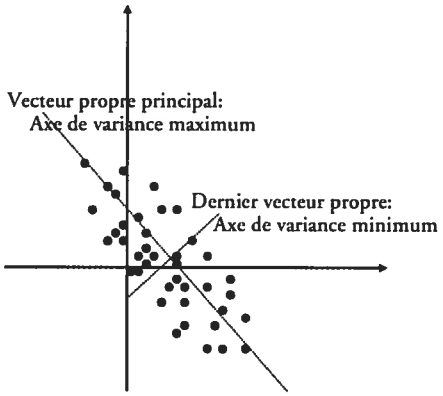

Figure 5.1 – *Sur cet exemple en deux dimensions, le lecteur reconnaîtra que la projection sur l'axe de variance maximum n'est autre qu'une régression linéaire, minimisant l'erreur quadratique de reconstruction. L'ACP n'est donc qu'une généralisation de ce principe à une projection en dimension K d'un point de dimension n.*

En bref, l'ACP projette les points dans un nouvel espace de variance maximale.

5.3 Généraliser ?

L'apprentissage non supervisé étant un des domaines les plus actifs de l'apprentissage automatique, il est nécessaire de s'attaquer à la question de la généralisation des algorithmes. Or, il va sans dire, dans le cadre non super-

visé, la notion de généralisation n'est plus nécessairement intuitive. Avec la croissance explosive des bases de données non supervisées accessibles par ordinateur et l'importance pressentie des techniques non supervisées dans les systèmes automatiques de demain, la nécessité d'un cadre théorique rigoureux se fait sentir plus que jamais. Entre autres questions, il est légitime de s'interroger sur la stabilité de ces algorithmes advenant de légères modifications des données. De même, on peut se demander comment formaliser l'extension de certains algorithmes, de partitionnement par exemple, à des données nouvelles.

Plusieurs travaux ont donc été faits en ce sens, quoique plusieurs questions soient toujours en suspend. Le lecteur peut ainsi se référer à (MIETZNER, OPPER et KINZEL 1994; WISKOTT et SEJNOWSKI 2002; ROTH, LANGE, BRAUN et BUHMANN 2002).

5.4 Apprentissage semi-supervisé

Dans plusieurs applications d'intérêt en apprentissage automatique — pensons entre autres à la reconnaissance de la parole, à la détection d'objets dans une image ou à la traduction de textes — on ne dispose que d'une faible proportion de données étiquetées relativement à la pléthore de données disponibles. Comme son nom l'indique, l'apprentissage semi-supervisé, à mi-chemin entre l'apprentissage supervisé et non-supervisé, est un cadre dans lequel on tente de combiner les deux sources d'information (BLUM et MITCHELL 1998) et d'en tirer profit.

Comme ces méthodes ne sont pas d'utilité dans ce mémoire, nous nous contenterons[‡] de dire que la majorité des algorithmes semi-supervisés reposent sur des hypothèses relatives à la distribution conjointe des entrées x et des sorties y, même si plusieurs tentent de se défaire de telles suppositions (SCHUURMANS

[‡]Pour une introduction au domaines et questions émergentes de l'apprentissage semi-supervisé, le lecteur peut se référer à (SEEGER 2001).

et SOUTHEY 2002). Pour les autres, deux grandes approches se partagent le plancher : paramétriques et non-paramétriques (BENGIO, DELALLEAU et LE ROUX 2004). Les premières (COZMAN, COHEN et CIRELO 2003) étant d'une inefficacité notoire lorsque la proportion de données non étiquetées est trop élevée, les secondes semblent vouloir s'imposer (SZUMMER et JAAKKOLA 2002; CHAPELLE, WESTON et SCHÖLKOPF 2003; BELKIN et NIYOGI 2003; ZHU, GHAHRAMANI et LAFFERTY 2003; ZHOU, BOUSQUET, NAVIN LAL, WESTON et SCHÖLKOPF 2004).

CHAPITRE 6

Fléau de la dimensionnalité

Dans moult applications d'intérêt, la dimension de l'espace des observations $\mathcal{X}$ peut être immense. Il suffit de penser à la reconnaissance d'image, par exemple. Si chaque observation est une image carrée en teintes de gris de 64 pixels de côté, une taille plus que modeste, déjà la représentation vectorielle de cette image est de dimension 4096 (64^2). En considérant qu'un algorithme d'apprentissage tente de modéliser, d'une façon ou d'une autre, la densité de laquelle sont tirées les observations, il y a fort à parier que les résultats ne seront que très peu valables si l'on attaque directement le vecteur de dimension 4096. En effet, à moins de disposer de plusieurs millions d'images, le nombre d'exemples sera bien trop faible face à la dimension de l'espace dans lequel on les retrouve. Ainsi, les données seront fort probablement dispersées de telle manière qu'il sera pratiquement impossible d'en déduire quoi que se soit (figure (6.1)).

6.1 Les fondements mathématiques du fléau

Pour se donner une intuition des fondements mathématiques du fléau, remarquons d'abord que que l'extrapolation est une tâche nettement plus complexe

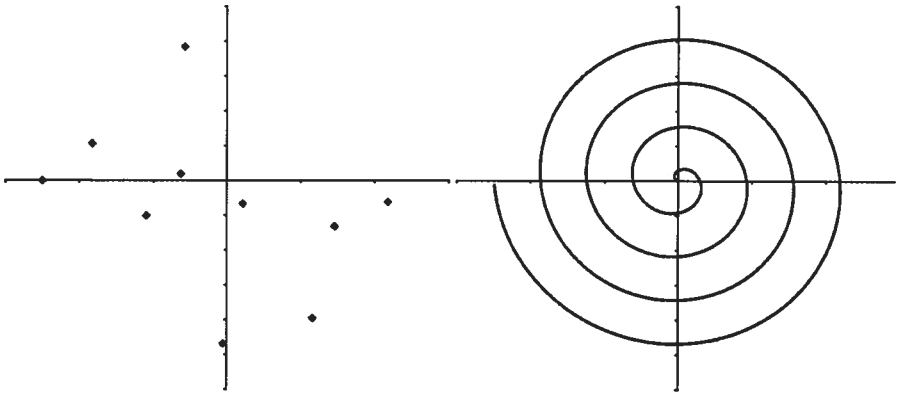

Figure 6.1 – *Dans la figure de gauche, on observe 10 points tirés de la variété représentée à droite. Ainsi, même en 2 dimensions avec 5 fois plus d'exemples que la taille de l'espace, force est d'admettre qu'aucun algorithme n'aurait pu approximer correctement la distribution. Imaginons maintenant une variété hautement non linéaire en dimension 4096.*

que l'interpolation. Rappelons-nous que dans le cadre de l'apprentissage supervisé, l'objectif est de minimiser l'erreur de généralisation

$$(\mathcal{L}f) = \int c(x, y, f(x))\ d\mu(x, y)$$

et que l'on choisit souvent, pour ce faire, de minimiser l'erreur empirique

$$(\mathcal{L}_{\mathcal{D}}f) = \frac{1}{m}\sum_{i=1}^{m} c(x_i, y_i, f(x_i))$$

qui convergerait, au sens de la loi des grands nombres (RICE 1994), vers l'erreur de généralisation si l'on disposait d'une infinité d'exemples indépendants. Le fait est que, plus souvent qu'autrement, les exemples disponibles le sont en nombre relativement restreint. En minimisant un critère comme l'erreur empirique (7.1), on n'optimise les paramètres que pour les régions de l'espace suffisamment peuplées par les observations. Ce faisant, il est relativement évident que le modèle ainsi obtenu n'aura qu'une piètre capacité de généralisation dans l'espace inexploré (figure 6.2).

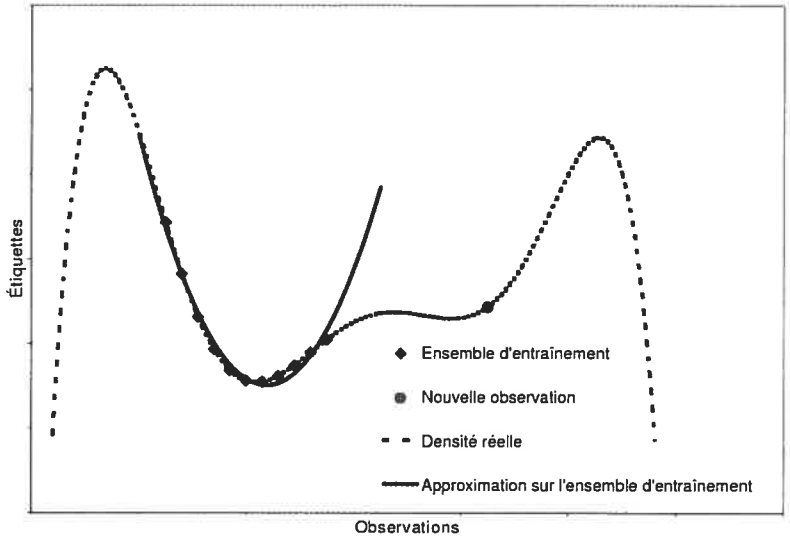

Figure 6.2 – *Sur les données d'entraînement, le modèle représenté par la ligne continue explique les données de façon plus que satisfaisante. Par contre, sa performance sur le nouvel exemple, hors de l'enveloppe convexe des points d'entraînement, est désastreuse. La tâche d'extrapolation est une tâche nettement plus ardue que la tâche d'interpolation.*

Maintenant, imaginons que les données soient tirées dans l'hypercube H de centre 0 et de volume unitaire en dimension n. D'abord, remarquons que D_m, l'enveloppe convexe des x_i, est de volume nul dès que $m \leq n$ puisque l'enveloppe convexe sera de dimension inférieure à celle de l'espace puisque, pour tout ensemble de n points (ou moins) dans $\mathbb{R}^n$, il est possible de trouver une variété de dimension au plus $(n-1)$ passant par lesdits n points. Ensuite, même si $m = n + 1$ il est possible de montrer que $Volume(D_m) \leq \frac{1}{n!}$ et tend donc très rapidement vers 0 lorsque n croît. Ainsi, la probabilité qu'une nouvelle observation se situe dans l'enveloppe convexe des exemples d'entraînement est pratiquement nulle et on se trouvera à extrapoler.

6.2 Fonction de sélection des caractéristiques

Une première solution peut provenir de l'extraction des caractéristiques les plus importantes par le biais de techniques de réduction de la dimensionnalité, comme nous l'avons vu précédemment.

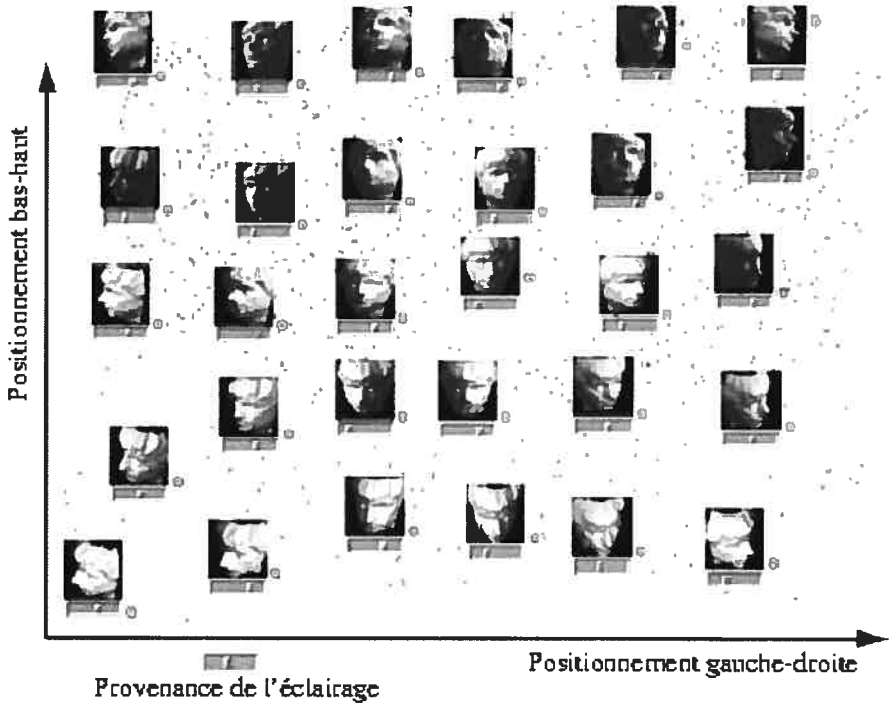

Figure 6.3 – *Dans cet exemple, en appliquant un algorithme de réduction de dimension (ici Isomap) sur les données en 4096 dimensions, on peut arriver à isoler 3 dimensions contenant la majorité de l'information.* (Source : (TENENBAUM, DE SILVA et LANGFORD 2000))

Or, non seulement ces techniques ne sont pas infaillibles, mais il arrive parfois que l'on préfère ajouter des caractéristiques plutôt que d'en retirer, les caractéristiques fondamentales n'étant pas toujours strictement incluses dans les caractéristiques fournies.

Définition 6.1 *Nous nommerons* **fonction de sélection des caractéristiques** *une application*

$$\begin{aligned} \varphi : \mathcal{X} &\to \mathcal{H} \\ x &\mapsto \varphi(x) \end{aligned}$$

où le codomaine $\mathcal{H}$ est appellé l'espace des caractéristiques.

Étant donnée une observation x de nature quelconque, cet outil nous permet d'obtenir une représentation vectorielle $\varphi(x)$ de l'observation en termes de caractéristiques qui peuvent décrire la nature et les particularités de x. Dans l'exemple des visages, on peut considérer que chaque exemple est d'abord une entité non numérique : une image en teintes de gris 64×64. Sur chaque image est d'abord appliqué une fonction de sélection des caractéristiques dont le domaine est l'espace des images en teinte de gris 64×64 et le codomaine l'espace $[0,1]^{4096}$, où 0 correspond à noir et 1 à blanc.

Pour obtenir les caractéristiques présentées à la figure (6.3), nous dirons que la fonction apprise par l'algorithme, f, est en fait une fonction de sélection des caractéristiques dont le domaine est $[0,1]^{4096}$ et le codomaine $[0,1]^3$, où 0 correspond à *gauche* ou *bas* et 1 à *droite* ou *haut*.

Cette dernière fonction de sélection des caractéristiques avait pour but d'appliquer à un exemple en haute dimension une description plus synthétique. Quoi qu'il en soit, l'application d'une fonction de sélection des caractéristiques pour augmenter la dimensionnalité peut s'avérer, dans certains cas, un outil d'une puissance indéniable.

Dans la figure (6.4), à gauche les points des deux classes sont clairement isolés par une ellipse, une fonction de décision quadratique. Or, en appliquant la fonction de sélection des caractéristiques non linéaire $\varphi([x]_1, [x]_2) = ([z]_1, [z]_2, [z]_3) = ([x]_1^2, [x]_2^2, [x]_1 [x]_2)$ l'ajout d'une dimension permet de séparer *linéairement* les exemples dans l'espace des caractéristiques (figure de droite), tâche impossible dans l'espace original.

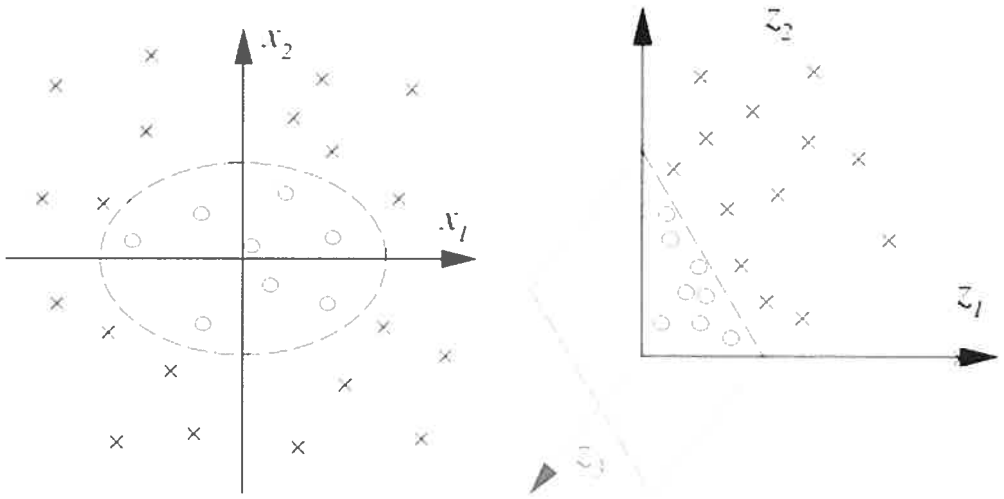

Figure 6.4 – *Fonction de sélection des caractéristiques sur un ellipse : Ajout d'une dimension;* $\varphi([x]_1, [x]_2) = ([z]_1, [z]_2, [z]_3) = ([x]_1^2, [x]_2^2, [x]_1 [x]_2)$. *(Source : (*SCHÖLKOPF *et* SMOLA *2002))*

De façon générale, remarquons que pour un nombre fini de points m, il existe toujours une application non linéaire en haute dimension qui permette de rendre les données linéairement séparables. Suite à cette dernière remarque, on se demandera pourquoi ne pas recourir systématiquement à une fonction de sélection des caractéristiques hautement non linéaire de dimension très élevée avant de procéder à l'apprentissage. Cette question, loin d'être saugrenue, trouve malheureusement pour première réponse, comme le lecteur l'aura sûrement pressenti, le fléau de la dimensionnalité.

CHAPITRE 7

Fonctions noyau

Les fonctions noyau, présentes dans la littérature depuis près d'un siècle (MERCER 1909), ne sont en fait qu'une mesure de similarité qui généralise le produit scalaire. Bien que quelques auteurs en aient fait usage sur le plan théorique à travers le siècle (AIZERMAN, BRAVERMAN et ROZONOÉR 1964; KOLMOGOROV 1941), il a fallu attendre le milieu des années 1990 pour que la communauté d'apprentissage automatique réalise la puissance pratique des fonctions noyau (BOSER, GUYON et VAPNIK 1992; SCHÖLKOPF, SMOLA et MULLER 1996; SCHÖLKOPF, SMOLA et MULLER 1998). Cette puissance, comme nous le verrons bientôt, est entre due au fait que les fonctions noyau permettent de contourner certains aspects du fléau de la dimensionalité.

7.1 Noyaux de Mercer

Introduisons maintenant une première définition de fonctions noyau.

Définition 7.1 *On appelle* **noyau de Mercer** — *ou tout simplement noyau* — *une fonction*

$$\mathrm{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \tag{7.1}$$

$$(x, x') \mapsto \mathrm{K}(x, x') \tag{7.2}$$

telle qu'il existe une fonction φ *et espace de Hilbert* $\mathcal{H}$ *muni d'un produit scalaire* $\langle ., . \rangle_{\mathcal{H}}$ *tels que*

$$\varphi : \mathcal{X} \to \mathcal{H} \tag{7.3}$$

$$x \mapsto \varphi(x) \tag{7.4}$$

et que

$$\mathrm{K}(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}} \tag{7.5}$$

Bref, en autant que $\mathrm{K}(x, x')$ se calcule en un nombre d'opérations raisonnable, φ peut projeter les observations dans un espace de dimension arbitrairement grand.

Une première question, des plus naturelles, consiste à déterminer quelles sont les fonctions qui répondent à cette définition. La première réponse revient au célèbre théorème de Mercer (MERCER 1909), d'où l'appellation *noyau de Mercer*.

Théorème 7.1 *Soit* $\mathrm{K} \in L_\infty^{\mathcal{X}^2}$, *une fonction symétrique (Définition (1.3)) à valeur réelle telle que*

$$T_k : L_2^{\mathcal{X}} \to L_2^{\mathcal{X}}$$

$$(T_k f)(x) := \int_{\mathcal{X}} \mathrm{K}(x, x') f(x') d\mu(x')$$

est défini positif; c'est-à-dire que pour tout $f \in L_2^{\mathcal{X}}$ *nous avons*

$$\int_{\mathcal{X}^2} \mathrm{K}(x, x') f(x) f(x') d\mu(x) d\mu(x') \geq 0$$

Posons $\psi_j \in L_2^{\mathcal{X}}$, $j = 1, \ldots, \mathcal{N}$ *les fonctions propres orthonormales de* T_k *et leurs valeurs propres associées* $\lambda_j > 0$ *(Définition (1.19)), en ordre croissant. Alors*

(i) $\lambda = (\lambda_1, \lambda_2, \ldots) \in l_1$ *(Définition (1.14))*

(ii) $\mathrm{K}(x, x') = \sum_{i=1}^{\mathcal{N}} \lambda_j \psi_j(x) \psi_j(x')$ *pour presque tout couple (x, x'). Si* $\mathcal{N}$ *tend vers l'infini, la série est absolument et uniformément convergente pour presque tout (x, x').*

Évidemment, en posant

$$\varphi : \mathcal{X} \to l_2^{\mathcal{N}} \tag{7.6}$$

$$x \mapsto (\sqrt{\lambda_j} \psi_j)_{j=1,\ldots,\mathcal{N}} \tag{7.7}$$

on obtient une expression de K comme produit scalaire (Équation (7.5)) dans l'espace de Hilbert $\mathcal{H} = l_2^{\mathcal{N}}$, où $\dim(\mathcal{H}) = \mathcal{N}$. Remarquons que rien n'empêche que $\mathcal{N}$ tende vers l'infini. Nous voilà donc munis d'une première caractérisation des fonctions noyau grâce au théorème de Mercer. Force est d'admettre cependant que cette caractérisation est un peu lourde.

7.2 Noyaux reproducteurs et applications sous-jacentes

Nous introduisons maintenant la notion de matrice de Gram, outil puissant pour l'analyse de fonctions noyau.

Définition 7.2 *Soit* $\mathrm{K} : \mathcal{X} \to \mathbb{R}$ *et soit l'ensemble* $\mathrm{S} = \{x_1, \ldots, x_m\} \subset \mathcal{X}$. *La matrice* $m \times m$ $\mathrm{M_K}$ *dont les éléments sont*

$$[\mathrm{M_K}]_{i,j} = \mathrm{K}(x_i, x_j)$$

est dite la **matrice de Gram** *de* K *sur* S.

Bref, la matrice de Gram permet de représenter concrètement l'application d'un noyau sur un ensemble de points. Nous entreprenons ici de faire le pont entre le théorème de Mercer et la matrice de Gram. La définition suivante permet un premier pas dans cette direction.

Définition 7.3 *Soit* $\mathcal{X}$ *un ensemble non vide. On appelle* **noyau défini positif** *une fonction* $\mathrm{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *telle que la matrice de Gram associée* $\mathrm{M_K}$ *est définie positive (Définition (1.16)) pour tout* $m \in \mathbb{N}$ *et pour tout ensemble* $\{x_1, \ldots, x_m\} \subset \mathcal{X}^m$.

Nous tenterons* maintenant d'établir qu'il existe une fonction Φ telle qu'un noyau défini positif K puisse être évalué par $\mathrm{K}(x, y) = \langle \Phi(x), \Phi(y) \rangle$. Pour ce faire, posons $\mathcal{X}$ un ensemble non vide et $\mathbb{R}^{\mathcal{X}} := \{f : \mathcal{X} \to \mathbb{R}\}$, l'ensemble des applications de $\mathcal{X}$ dans $\mathbb{R}$. Étant donné K un noyau défini positif sur $\mathcal{X}$, définissons

$$\Phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}} \tag{7.8}$$

$$x \mapsto \mathrm{K}(x, .). \tag{7.9}$$

Chaque observation x engendre ainsi une fonction sur $\mathcal{X}$. Dans cette optique, une observation peut être représentée par sa mesure de similarité avec tout autre point de $\mathcal{X}$. Non seulement cette représentation est-elle riche en information, mais nous prétendons qu'elle est aussi implicite au noyau K. Pour s'en

*Avouons que nous sommes, dans cette tentative, très inspirés par (SCHÖLKOPF et SMOLA 2002).

assurer, il est possible de construire un espace vectoriel $\mathcal{H}$ tel que

$$f \in \mathcal{H} \Leftrightarrow f : \mathcal{X} \to \mathbb{R} \text{ et } \exists m \in \mathbb{N}, \mathcal{D} \in \mathcal{X}^m, \alpha \in \mathbb{R}^m$$
$$\text{tels que } f(x) = \sum_{i=0}^{m} [\alpha]_i \, \mathrm{K}(x_i, x), x_i \in \mathcal{D} \tag{7.10}$$

sur lequel on définit le produit scalaire

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i=0}^{m} \sum_{j=0}^{m'} [\alpha]_i \, [\alpha']_j \, \mathrm{K}(x_i, x'_j), \tag{7.11}$$

où $x_i \in \mathcal{D}$ et $x'_j \in \mathcal{D}'$, $\mathcal{D}$ et $\mathcal{D}'$ étant les ensembles définissant respectivement f et g. Il est aisé de démontrer que $\langle f, g \rangle_{\mathcal{H}}$ est bien défini et est bel et bien un produit scalaire (Schölkopf et Smola 2002).

Remarquons que $\Phi(x)$ est dans $\mathcal{H}$; il suffit de prendre

$$m' = 1, \mathcal{D}' = \{x\}, \alpha' = 1. \tag{7.12}$$

Son produit scalaire avec toute fonction $f \in \mathcal{H}$ est donc

$$\langle \Phi(x), f \rangle = \langle \mathrm{K}(., x), f \rangle \tag{7.13}$$
$$= \sum_{i=0}^{m} [\alpha]_i \, \alpha' \, \mathrm{K}(x_i, x) \tag{7.14}$$
$$= f(x), \tag{7.15}$$

où l'utilisation (7.12) dans (7.11) entraîne le passage de l'équation (7.13) à (7.14). L'équation (7.15) montre K sous l'angle de la **fonction d'évaluation**. Afin de démontrer que la représentation de x par $\Phi(x)$ est implicite au noyau K il suffit d'observer le cas particulier de (7.15) où $f = \Phi(x')$

$$\langle \Phi(x), \Phi(x') \rangle = \mathrm{K}(x, x'). \tag{7.16}$$

Les équations (7.15) et (7.16) ont valu aux noyaux définis positifs l'appellation de noyaux reproducteurs (BERG, CHRISTENSEN et RESSEL 1984; SAITOH 1988).

7.3 Espaces de Hilbert des noyaux reproducteurs

Dans la section précédente nous avons introduit avec les équations (7.10) et (7.11) un espace $\mathcal{H}$ pré-Hilbertien (Définition (1.8)). Transformer un espace pré-Hilbertien en espace de Hilbert est relativement simple et permet, entre autres, de s'assurer que toute projection est bien définie. Pour procéder à ladite transformation, il suffit de considérer la norme $\|f\| = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$ et d'ajouter à $\mathcal{H}$ les points de convergence au sens de cette norme.

Définition 7.4 (SCHÖLKOPF et SMOLA 2002) *Soit $\mathcal{X}$ un ensemble non vide et $\mathcal{H}$ un espace de Hilbert de fonctions $f : \mathcal{X} \to \mathbb{R}$. Alors $\mathcal{H}$ est l'espace de Hilbert d'un noyau reproducteur s'il existe un noyau $\mathrm{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ tel que*

(i) $\langle f, \mathrm{K}(x, .) \rangle = f(x), \forall f \in \mathcal{H}$

(ii) $\mathcal{H} = \overline{\mathrm{span}\{\mathrm{K}(x, .) \,|\, x \in \mathcal{X}\}}$, *où $\overline{E}$ dénote la fermeture de l'ensemble E (?).*

Enfin, notons que l'équation (7.16) nous permet aussi d'affirmer que les noyaux définis positifs sont aussi des noyaux de Mercer. Si l'on rappelle le résultat suivant,

Théorème 7.2 *(BERG, CHRISTENSEN et RESSEL 1984) Soit φ une fonction quelconque d'un compact $\mathcal{X}$ dans $\mathbb{C}$, alors $\mathrm{K}(x, x') = \langle \varphi(x), \varphi(x') \rangle$ est une fonction définie positive.*

qui n'est, à toutes fins pratiques, que la réciproque du théorème de Mercer, on remarque que les noyaux de Mercer sont forcément définis positifs.

le produit scalaire respecte trivialement la définition (7.1). Ainsi, l'utilisation de différents noyaux en lieu de produit scalaire permettra la généralisation en non linéaire de nombre d'algorithmes linéaires. C'est cette substitution que l'on qualifie d'**astuce du noyau**.

Dans les dernières années, une quantité impressionnante d'algorithmes supervisés et non supervisés ont été développés en utilisant cette astuce. (Bengio, Delalleau, Le Roux, Paiement, Vincent et Ouimet 2004) constitue non seulement une bonne introduction à une vaste gamme de ces méthodes — techniques de réduction de dimension telles MDS (Torgerson 1958; Gower 1966), Isomap (Tenenbaum, de Silva et Langford 2000), LLE (Roweis et Saul 2000), etc. — mais les auteurs s'inspirent aussi de liens que font (Williams et Seeger 2000) entre la projection faite par l'ACP à noyaux (Schölkopf, Smola et Muller 1999) et la formule de Nyström (Baker 1977) pour développer un cadre théorique permettant de regrouper et d'étendre lesdites techniques.

7.5 Régularisation et noyaux : Théorème du représentant

À la section 4.4.1, nous avons introduit le concept de régularisation, concept d'une puissance indéniable dans la poursuite de l'objectif de généralisation. Dans cette section, nous introduisons un théorème central dans les applications pratiques des méthodes à noyaux. En effet, dans moult modèles d'estimations statistiques, le **théorème du représentant**[†] sert de justification théorique à l'utilisation pratiques des fonctions noyau.

Théorème 7.3 *Soient* $\Omega : \mathbb{R}^+ \to \mathbb{R}$ *une fonction strictement croissante,* $\mathcal{X}$ *un ensemble non vide et* $(\mathcal{L}_D f) : (\mathcal{X} \times \mathbb{R}^2)^m \to \mathbb{R} \bigcup \{\infty\}$ *une erreur empirique*

[†]En anglais : *Representer Theorem.*

quelconque. Alors, étant donnés un noyau reproducteur K *et l'espace de Hilbert sous-jacent* $\mathcal{H}$*, chaque minimum local* $f \in \mathcal{H}$ *de l'erreur régularisée*

$$(\mathcal{L}^{\Omega}_{\mathcal{D}} f) = (\mathcal{L}_{\mathcal{D}} f)((x_1, y_1, f(x_1)), \ldots, (x_m, y_m, f(x_m))) + \Omega(\ \|f\|^2_{\mathcal{H}}\ ) \qquad (7.17)$$

admet une représentation de la forme

$$f(x) = \sum_{i=1}^{m} \alpha_i \, \mathrm{K}(x_i, x) \qquad (7.18)$$

Évidemment, pour assurer l'unicité du minimum, on doit exiger la convexité de $(\mathcal{L}^{\Omega}_{\mathcal{D}} f)$.

Prouvé d'abord exclusivement pour les moindres carrés par (Kimeldorf et Wahba 1971), ce n'est que près de 20 ans plus tard que le théorème fût généralisé par (Cox et O'Sullivan 1990) pour admettre une fonction de coût $c : (X \times \mathbb{R}^2) \to \mathbb{R}$ quelconque. La version présentée ici, qui permet une erreur empirique combinant les exemples, a été introduite par (Schölkopf, Herbrich, Smola et Williamson 2001) et on en trouve une élégante preuve dans (Schölkopf et Smola 2002) que nous répétons ici.

Preuve Pour toute fonction $f \in \mathcal{H}$, il existe une décomposition de f en deux parties, l'une, $f_{\|}$, dans l'espace engendré par les $\mathrm{K}(x_i, .)_{x_i \in \mathcal{X}_{\mathcal{D}}}$ et l'autre, $f_{\perp}$, dans son complément de telle manière que

$$f(x) = f_{\|}(x) + f_{\perp}(x) = \sum_{i=1}^{m} \alpha_i \, \mathrm{K}(x_i, x) + f_{\perp}(x) \qquad (7.19)$$

où $\alpha_i \in \mathbb{R}$ et où $f_\perp \in \mathcal{H}$ est telle que $\langle f_\perp(.), \mathrm{K}(x_i, .)\rangle_{\mathcal{H}} = 0, \forall i \in \{1, \ldots, m\}$. En regard de l'équation (7.15), on peut aussi écrire

$$f(x_j) = \langle f(.), \mathrm{K}(x_j, .)\rangle \tag{7.20}$$

$$= \sum_{i=1}^{m} \alpha_i \mathrm{K}(x_i, x_j) + \langle f_\perp(.), \mathrm{K}(x_j, .)\rangle_{\mathcal{H}} \tag{7.21}$$

$$= \sum_{i=1}^{m} \alpha_i \mathrm{K}(x_i, x_j), \forall j \in \{1, \ldots, m\} \tag{7.22}$$

De plus, $\forall f_\perp$,

$$\Omega(\|f\|_{\mathcal{H}}^2) = \Omega\left(\left\|\sum_{i=1}^{m} \alpha_i \mathrm{K}(x_i, .)\right\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2\right) \tag{7.23}$$

$$\geq \Omega\left(\left\|\sum_{i=1}^{m} \alpha_i \mathrm{K}(x_i, .)\right\|_{\mathcal{H}}^2\right) \tag{7.24}$$

Ainsi, $\forall \alpha \in \mathbb{R}^m$, $(\mathcal{L}_{\mathcal{D}}^{\Omega} f)$ est minimisée pour $f_\perp = 0$. ■

Partie IV

Notre modèle

L'apprentissage automatique est connu pour être un outil fiable et puissant pour la gestion de portefeuilles boursiers (Weigend, Abu-Mostafa et Refenes 1997). Dans la plupart des cas, on utilise des réseaux de neurones pour effectuer une régression sous le coût quadratique afin d'obtenir des prédictions sur les séries de rendements. La gestion de portefeuille est ensuite déléguée à un modèle économique classique.

Depuis quelques années, certains résultats autant théoriques que pratiques tendent à démontrer que, plutôt que de reléguer le modèle statistique au rôle d'intermédiaire, on gagne à optimiser un modèle sous un critère financier afin d'apprendre directement les positions à prendre sur le marché (Bengio 1997; Chapados 2000).

Avec le développement des méthodes à noyaux, certains auteurs ont cherché à remplacer les réseaux de neurones par des modèles à noyaux dans la prédiction de séries financières (Kivinen, Smola et Williamson 2002). Toutefois, on ne confine encore les modèles qu'à un rôle de prédiction, laissant la décision pour la théorie économique.

Nous suggérons donc un modèle à noyaux qui assume la prise de décision. Nous ne supposons aucun oracle fournissant les suites optimales de positions à prendre, sortant ainsi légèrement du cadre de régression classique pour se rapprocher des méthodes de maximisation de la vraisemblance.

CHAPITRE 8

Formalisation de la problématique

Maintenant les théories de l'économie et de l'apprentissage automatique sommairement exposées, nous somme près à attaquer la problématique énoncée en introduction (**Partie I**).

Rappelons que nous considérons ici un problème de gestion d'un portefeuille boursier. Par souci de simplicité, nous considérons un modèle à temps discret, c'est-à-dire dans lequel une **période** (une journée ou un mois, par exemple) s'écoule entre les temps t et $t+1$, $t \in \mathbb{N}$. Nous considérons la période t comme étant celle qui s'écoule entre les temps $t-1$ et t.

8.1 L'état du système

Dans ce modèle, nous segmenterons l'état du système au temps t en deux composantes distinctes ; soient (i) l'information exogène produite par le marché et (ii) notre position sur le marché.

8.1.1 L'information exogène produite par le marché

Les observations

L'observation au temps t est $x_t \in \mathbb{R}^n$, un résumé de l'information générée par le marché jusqu'au temps t. Notons que ces observations sont intrinsèquement non-stationnaires, c'est-à-dire que la loi de probabilité gouvernant l'évolution du marché dépend du temps, formellement, il existe δ tel que

$$\Pr(x_{t_1}, x_{t_2}, \ldots, x_{t_s}) \neq \Pr(x_{t_1+\delta}, x_{t_2+\delta}, \ldots, x_{t_s+\delta}) \tag{8.1}$$

Conséquemment, nous noterons $\mu_t(x_t \mid x_\tau, \tau = 1, \ldots, t-1)$ la probabilité d'observer x_t au temps t étant donné le passé.

L'horizon

Un concept intéressant lorsque l'on travaille avec des séries chronologiques est celui d'horizon. L'horizon, $h \geq 1$, est le nombre de périodes séparant la prise de décision (la prédiction) du moment auquel on peut évaluer la qualité de cette prédiction.

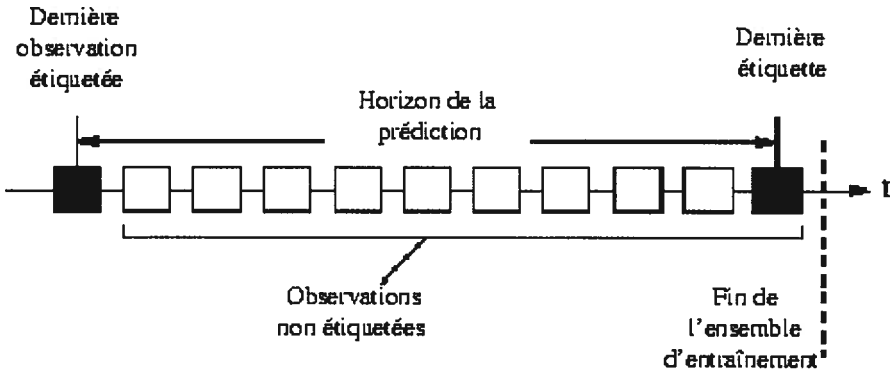

Figure 8.1 – *Horizon* (Source : (CHAPADOS et BENGIO 2003))

À la limite, h sera égal à 1 : la décision prise au temps t sera appliqué immédiatement avant le temps $t+1$.

Les étiquettes

Dans ce modèle, nous considérons l'étiquette au temps t, y_{t+h}, comme étant l'information disponible via les les agents du marché au temps $t+h$. Quoique $y_{t+h} = x_{t+h}$, nous conserverons la notation y_{t+h} pour bien distinguer les étiquettes des observations.

Évidemment, il n'est pas question de prédire toute l'information que générera le marché entre t et $t+h$. Toutefois, comme nous l'avons mentionné précédemment, une des particularités de notre formulation est que nous ne supposons aucun oracle qui puisse fournir le portefeuille optimal sur le marché. Nous allons plutôt définir (§ 9.1) une erreur empirique qui évalue, compte tenu des rendements calculés à partir de y_{t+h}, le portefeuille proposé connaissant x_t.

Ensemble d'entraînement et paire de test

Nous segmenterons l'information disponible au temps t en l'ensemble d'entraînement au temps t

$$\mathcal{D}_t = \{(x_{s-h}, y_s)\}_{s=t-(\Delta+h-1)}^{t-h} \tag{8.2}$$

et la paire de test $\mathcal{T}_t = \{(x_{t-h}, y_t)\}$.

Deux raisons motivent le choix de $\mathcal{D}_t$ comme il est ici présenté. D'abord, la dernière paire d'entraînement doit impérativement être (x_{t-2h}, y_{t-h}). En effet, comme nous désirons tester avec (x_{t-h}, y_t), l'ensemble d'entraînement ne peut contenir d'étiquette y_τ, $\tau > t-h$, puisque cette étiquette se révélerait être future à l'observation de test. Ensuite, nous désirons utiliser une fenêtre Δ

restreignant l'étendue du passé considéré et, de ce fait, limitant la taille de l'ensemble d'entraînement à Δ paires d'entraînement*.

8.1.2 Notre position sur le marché

Définition 8.1 *Un* **portefeuille** $\mathbf{w}_t$ *défini sur un ensemble de* N *biens est le vecteur des quantités possédées de chaque bien au temps* t *donné :*

$$\mathbf{w}_t = (w_{1t}, w_{2t}, \ldots, w_{Nt})',$$

où $w_{kt} \in \mathbb{R}$.

Il est important de remarquer qu'ici les poids sont absolus et non relatifs comme dans la **Partie II**. Cette notation permet d'éliminer la contrainte de somme à 1 qui émerge lors d'une optimisation avec des poids relatifs. D'autre part, si l'on note les poids relatifs utilisés dans la **Partie II** par $w^{\%}_{kt}$, on peut aisément les retrouver en partant des poids absolus w_{kt} utilisés ici :

$$w^{\%}_{kt} = \frac{|w_{kt}|p_{kt}}{\sum_{k'=1}^{N} |w_{k't}|p_{k't}} \tag{8.3}$$

où p_{kt} est le prix de chaque unité du bien k au temps t. Remarquons l'usage de la valeur absolue des poids w_{kt}. En effet, sur le marché des contrats à terme boursiers, rien n'empêche la vente d'un contrat que l'on ne possède pas. Ce faisant, la position dans ce bien est considérée négative. Ce mécanisme, appelé **vente à découvert**†, requiert toutefois que l'investisseur porte à sa marge (§ 2.4.2) le même montant qu'il y aurait porté si sa position avait été positive, d'où la valeur absolue.

*Nous serons plus à même d'expliquer pourquoi à la section 9.3

† En anglais : *Short Sale.*

8.2 La décision

L'enjeu fondamental du modèle que nous suggérerons (Chapitre 9) est essentiellement la prise de décision, à savoir la composition d'un portefeuille futur étant donné l'historique présent. Nous laisserons donc le détail de la prise de décision pour le chapitre suivant et n'en élaborerons ici que l'intuition.

Étant donné l'ensemble d'entraînement au temps t, ne contenant que les observations jusqu'à $(t-2h)$, nous désirons apprendre une règle de décision $f(.)$ permettant d'établir le portefeuille au temps t à l'aide de l'observation à $(t-h)$, c'est à dire

$$w_{kt} = f_k(x_{t-h}) \tag{8.4}$$

$$f = (f_1, \ldots, f_N) \tag{8.5}$$

Or, il y a fort à parier que les f_k seront hautement non linéaires. Plutôt que de chercher à faire directement une régression non linéaire dans l'espace original, une approche désormais classique est de projeter les observations dans un espace de caractéristiques à l'aide d'une fonction hautement non linéaire ϕ et de procéder à une régression linéaire dans le nouvel espace. Formellement, on se dote de $\phi : \mathcal{X} \to \mathcal{H}$ et on recherche $\boldsymbol{\beta}_k^{(t)}$ tel qu'au temps t

$$\begin{aligned} f_k(x_{t-h}) &= f_k(x_{t-h}; \boldsymbol{\beta}_k^{(t)}) \\ &= \left\langle \boldsymbol{\beta}_k^{(t)}, \phi(x_{t-h}) \right\rangle_{\mathcal{H}} \end{aligned} \tag{8.6}$$

Rappelons que notre objectif est de maximiser le compromis moyenne-variance du rendement mensuel moyen de notre stratégie de gestion de portefeuille. Le critère de l'apprentissage permettant d'obtenir la règle de décision f, que nous développerons à la section 9.1, devra être pensé en conséquence.

8.3 La fonction de transition

Sous les conditions énoncées précédemment, la fonction de transition, définissant l'évolution du système à travers le temps, est toute simple :

$$\mathbf{w}_t = f(x_{t-h}; \boldsymbol{\beta}^{(t)}) \tag{8.7}$$

$$x_t \sim \mu_t(. \mid x_{t-1}) \tag{8.8}$$

où μ un processus inconnu gouvernant l'évolution du marché et où $\boldsymbol{\beta}^{(t)}$ la concaténation des $\boldsymbol{\beta}_k^{(t)}$, $k = 1, \ldots, N$. À la lecture de la fonction de transition,

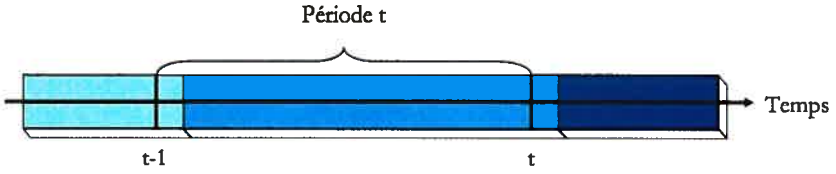

Figure 8.2 – *Une Période : Prenons l'exemple des marchés de contrats à terme boursiers; à la fermeture des marchés à 15h30 (16h30 heure de Chicago), nous considérons la journée $(t-1)$ comme étant terminée. Nous avons à notre disposition toutes les informations exogènes produites par les marchés à ce jour. Compte tenu du prix à la fermeture, nous sommes en mesure de mesurer notre performance passée et d'entraîner le modèle pour mettre à jour f à travers θ^t. Quoi qu'il en soit nous ne serons pas en mesure d'appliquer la nouvelle règle de décision avant l'ouverture des marchés le lendemain matin (début de la journée t).*

Un investisseur expérimenté nous ferait aussi remarquer qu'il existe des mécanismes d'échange hors des heures d'ouverture et que les prix peuvent donc évoluer, à notre insu, entre la fermeture de la journée t et l'ouverture de la journée $t+1$ (d'où le décalage entre les couleurs et les périodes sur le schéma). Ainsi, nous entraînons notre modèle avec une information imparfaite. Toutes ces technicalités devront être prises en compte dans certains modules d'exécution (§ 10.3.1).

le lecteur aura peut-être remarqué que nous ne fournissons pas le portefeuille actuel au modèle pour sa prise de décision. Nous considérons donc que le modèle trouve le portefeuille risqué idéal compte tenu des données sur les biens, mais sans égard au portefeuille actuel. Nous reviendrons à la section (§ 11.2) sur les conséquences de cette formulation.

CHAPITRE 9

Le modèle

En nous basant sur la précédente formalisation, nous sommes maintenant en mesure de présenter le modèle à proprement parler afin d'en venir à un problème d'optimisation complet.

9.1 Minimisation de l'erreur empirique

D'une importance capitale quant à la régression obtenue, l'erreur empirique est une des pièces fondamentales d'un modèle. Nous proposons ici l'erreur empirique, inspirée de principes économiques présentés à la **Partie II**, qui gouvernera l'apprentissage de notre règle de décision (§ 8.2).

Au chapitre 3.2, nous avons mentionné à moult reprises l'importance d'obtenir un bon compromis moyenne-variance. Conséquemment, nous avons introduit une fonction d'utilité

$$\mathrm{U}_P = \mathrm{E}[R_P] - \frac{1}{2}\mathrm{A}\sigma_P^2 \tag{9.1}$$

qui, pour $\mathrm{A} > 0$, est proportionnelle au rendement espéré et inversement proportionnelle au risque encouru. Le précédente propriété laisse croire qu'une

optimisation des paramètres entraînant le choix du portefeuille sous ce critère d'utilité nous permettra d'obtenir un portefeuille décemment construit.

Notons que nous ne supposons pas connaître le processus gouvernant l'évolution du marché. Ainsi, nous remplaçons U_P par l'estimateur

$$\hat{\mathrm{U}}_P = \overline{r}_P - \frac{A}{2}\hat{\sigma}_P^2 \tag{9.2}$$

où

$$S_l := \sum_{i=1}^{m} r_{P_i}^l, \tag{9.3}$$

$$\overline{r}_P := \frac{S_l}{m}, \tag{9.4}$$

$$\hat{\sigma}_P^2 := \frac{mS_2 - S_1^2}{m(m-1)}. \tag{9.5}$$

En somme, à chaque période, notre modèle minimisera l'erreur empirique suivante :

$$(\mathcal{L}_{\mathcal{D}_t} f)(\{x_{s-h}, y_s, f(x_{s-h})\}_{s=t-(\Delta+h-1)}^{t-h}) = -\overline{r}_P(\boldsymbol{\beta}^{(t)}) + \frac{A}{2}\hat{\sigma}_P^2(\boldsymbol{\beta}^{(t)}) \tag{9.6}$$

$$\overline{r}_P(\boldsymbol{\beta}^{(t)}) = \frac{1}{\Delta}\sum_{s=t-(\Delta+h-1)}^{t-h} r_{P_s}(\boldsymbol{\beta}^{(t)}) \tag{9.7}$$

$$\hat{\sigma}_P^2(\boldsymbol{\beta}^{(t)}) = \frac{\Delta\sum_{s=t-(\Delta+h-1)}^{t-h} r_{P_s}(\boldsymbol{\beta}^{(t)}) + \sum_{s=t-(\Delta+h-1)}^{t-h} r_{P_s}^2(\boldsymbol{\beta}^{(t)})}{\Delta(\Delta-1)} \tag{9.8}$$

$$r_{P_s}(\boldsymbol{\beta}^{(t)}) = \sum_{k=1}^{N} \left\langle \boldsymbol{\beta}_k^{(t)}, \phi(x_{s-h}) \right\rangle_{\mathcal{H}} r_{ki}(y_s) \tag{9.9}$$

Notons que la valeur de A devra éventuellement être judicieusement choisie.

9.2 Absence de normalisation

Ayant accès à tout l'historique et n'étant évalué que sur la base des rendements, le modèle pourrait avoir tendance à faire croître (resp. décroître) démesurément les positions à prendre dans les biens ayant un rendement positif (resp. négatif) à un moment donné. En somme, les portefeuilles optimaux pourraient ressembler à $\mathbf{w}_t = (+\infty, +\infty, -\infty, \ldots, -\infty)$, ce qui n'est évidemment pas très intéressant. Or, le même genre de problème est souvent rencontré dans l'optimisation de réseaux neuronaux et une solution classique (HINTON 1986) est la pénalisation des poids* qui consiste à ajouter à l'erreur empirique un terme proportionnel à $\sum_i w_i^2$.

Nous utiliserons donc une fonction de pénalité analogue, mais où nous pondérerons les positions par les prix respectifs des biens†, c'est à dire

$$\sum_{s=t-(\Delta+h-1)}^{t-h} \sum_{k=0}^{N-1} (w_{k\tau} p_{k\tau})^2 = \sum_{s=t-(\Delta+h-1)}^{t-h} \sum_{k=0}^{N-1} \left(\left\langle \boldsymbol{\beta}_k^{(t)}, \phi(x_{s-h}) \right\rangle_{\mathcal{H}} p_{k\tau} \right)^2 \tag{9.10}$$

$$= \sum_{k=0}^{N-1} p_{k\tau}^2 \boldsymbol{\beta}_k^{(t)^{\mathrm{T}}} \sum_{s,s'=t-(\Delta+h-1)}^{t-h} \phi(x_{s-h}) \phi(x_{s'-h})^{\mathrm{T}} \boldsymbol{\beta}_k^{(t)} \tag{9.11}$$

$$= \sum_{k=0}^{N-1} (p_{k\tau} \boldsymbol{\beta}_k^{(t)})^{\mathrm{T}} \mathrm{C} (p_{k\tau} \boldsymbol{\beta}_k^{(t)}) \tag{9.12}$$

où C est définie positive (Définition (1.16)) et $p_{kt} > 0, \forall k, t$. Il est donc possible de considérer une norme $\|.\|_{p\mathrm{C}}$ sur $f = (f_1, \ldots, f_N)$ définie suivant l'équation (9.12), c'est à dire

$$\|f\|_{p\mathrm{C}} := \sum_{k=0}^{N-1} (p_{k\tau} \boldsymbol{\beta}_k^{(t)})^{\mathrm{T}} \mathrm{C} (p_{k\tau} \boldsymbol{\beta}_k^{(t)}). \tag{9.13}$$

*En anglais : *weight decay.*

† La pondération par le prix est évidemment nécessaire en regard de l'équation (8.3).

Bref, nous considérerons la fonction de régularisation suivante :

$$\Omega(\|f\|_{pC}) = \gamma \|f\|_{pC} \tag{9.14}$$

où γ est un paramètre de régularisation qu'il faudra éventuellement déterminer.

9.3 Version vectorielle du théorème du représentant

Théorème 9.1 *Soient* $\Omega : \mathbb{R}_+ \to \mathbb{R}$ *une fonction croissante,* $\mathcal{X}$ *et* $\mathcal{Y}$ *des ensembles non-vides et*

$$(\mathcal{L}_{\mathcal{D}} f) : (\mathcal{X} \times \mathcal{Y} \times \mathbb{R}^N)^m \to \mathbb{R} \bigcup \{\infty\} \tag{9.15}$$

une erreur empirique quelconque. Enfin, étant donnés un noyau reproducteur K *et l'espace de Hilbert sous-jacent* $\mathcal{H}$*, posons*

$$\mathcal{H}^N = \{f = (f_1, \ldots, f_N) \,|\, f_k \in \mathcal{H}, \forall k\} \tag{9.16}$$

muni du produit scalaire $\langle f, g \rangle_{\mathcal{H}^N} = \sum_{k=1}^{N} \langle f, g \rangle_{\mathcal{H}}$.

Alors chaque minimum local $f \in \mathcal{H}^N$ *de l'erreur régularisée*

$$(\mathcal{L}^{\Omega}_{\mathcal{D}} f) = (\mathcal{L}_{\mathcal{D}} f)(x_1, y_1, f(x_1), \ldots, \quad x_m, y_m, f(x_m)) + \Omega(\, \|f\|^2_{\mathcal{H}^N}\,) \tag{9.17}$$

admet une représentation telle que, pour tout k*,*

$$f_k(x) = \sum_{i=1}^{m} \alpha_{ki} \, \mathrm{K}(x_i, x). \tag{9.18}$$

Preuve D'abord, en analysant la preuve du théorème du représentant (Théorème 7.3) de plus près, nous remarquons que le théorème aurait très bien pu

être énoncé pour $(\mathcal{L}_{\mathcal{D}} f) : (\mathcal{X} \times \mathcal{Y} \times \mathbb{R})^m \to \mathbb{R} \bigcup \{\infty\}$ sans que la preuve n'en soit affectée. Les auteurs (SCHÖLKOPF, HERBRICH, SMOLA et WILLIAMSON 2001) ont probablement eu recours à $\mathcal{Y} = \mathbb{R}$ dans l'optique où, dans la quasi-totalité des cas de régression, $f(x) \in \mathcal{Y}$.

Maintenant, remarquons que, pour toute fonction $f \in \mathcal{H}^N$, il existe une décomposition de f en deux parties. La première étant de la forme

$$f^{\|} = (f_1^{\|}, \ldots, f_N^{\|}) \tag{9.19}$$

où les $f_k^{\|}$ sont dans l'espace engendré par les $\mathrm{K}(x_i, .)_{x_i \in \mathcal{X}_{\mathcal{D}}}$ et la seconde,

$$f^{\perp} = (f_1^{\perp}, \ldots, f_N^{\perp}), \tag{9.20}$$

telle que les $f_k^{\perp}$ soient dans le complément dudit espace.

Ainsi,

$$f(x) = f^{\|}(x) + f^{\perp}(x) \tag{9.21}$$

$$= (\sum_{i=1}^{m} \alpha_{1i} \mathrm{K}(x_i, x) + f_1^{\perp}(x), \ldots, \sum_{i=1}^{m} \alpha_{Ni} \mathrm{K}(x_i, x) + f_N^{\perp}(x)) \tag{9.22}$$

où $\alpha_{ki} \in \mathbb{R}$ et où $f_k^{\perp} \in \mathcal{H}$ est telle que $\left\langle f_k^{\perp}(.), \mathrm{K}(x_i, .) \right\rangle_{\mathcal{H}} = 0, \forall (k, i) \in \{1, \ldots, N\} \times \{1, \ldots, m\}$. En regard de l'équation (7.15), on peut aussi écrire

$$f_k(x_j) = \left\langle f_k(.), \mathrm{K}(x_j, .) \right\rangle \tag{9.23}$$

$$= \sum_{i=1}^{m} \alpha_{ki} \mathrm{K}(x_i, x_j) + \left\langle f_k^{\perp}(.), \mathrm{K}(x_j, .) \right\rangle_{\mathcal{H}} \tag{9.24}$$

$$= \sum_{i=1}^{m} \alpha_{ki} \mathrm{K}(x_i, x_j), \forall j \in \{1, \ldots, m\} \tag{9.25}$$

De plus, $\forall f^{\perp}$,

$$\Omega(\|f\|^2_{\mathcal{H}^N}) = \Omega\left(\sum_{k=1}^{N}\left[\left\|\sum_{i=1}^{m}\alpha_{ki}\,\mathrm{K}(x_i,.)\right\|^2_{\mathcal{H}} + \left\|f_k^{\perp}\right\|^2_{\mathcal{H}}\right]\right) \tag{9.26}$$

$$\geq \Omega\left(\sum_{k=1}^{N}\left\|\sum_{i=1}^{m}\alpha_{ki}\,\mathrm{K}(x_i,.)\right\|^2_{\mathcal{H}}\right). \tag{9.27}$$

Ainsi, $(\mathcal{L}^{\Omega}_{\mathcal{D}}f)$ est minimisée pour $f^{\perp}=0$. Remarquons que les équations (9.26) et (9.27) sont valides pour toute norme sur $\mathcal{H}^N$ ■

À notre connaissance, cette forme du théorème du représentant est nouvelle. Or, dans notre problème, cette formulation est la bienvenue. En effet, si nous exprimons notre fonction de sélection des caractéristiques ϕ comme étant du type

$$\begin{aligned}\phi : \mathcal{X} &\to \mathbb{R}^{\mathcal{X}}\\ x &\mapsto \mathrm{K}(x,.)\end{aligned}$$

où K est un noyau reproducteur, il découle du théorème précédent que tout $\boldsymbol{\beta}_k^{(t)}$ se trouve dans l'espace engendré par les $\{\Phi(x_i)\}_{x_i\in\mathcal{X}_{\mathcal{D}}}$, c'est à dire que l'on peut réécrire l'équation (8.6) comme suit :

$$f_k(x_{t-h};\boldsymbol{\beta}_k^{(t)}) = \left\langle \boldsymbol{\beta}_k^{(t)}, \phi(x_{t-h})\right\rangle_{\mathcal{H}} \tag{9.28}$$

$$= \left\langle \sum_{s=t-(\Delta-h+1)}^{t-h} \alpha_{ks}^{(t)}\phi(x_{s-h}), \phi(x_{t-h})\right\rangle_{\mathcal{H}} \tag{9.29}$$

$$=: f_k(x_{t-h};\boldsymbol{\alpha}_k^{(t)}) \tag{9.30}$$

où $\boldsymbol{\alpha}_k^{(t)}$ est le vecteur de paramètres associés au bien k au temps t. Aussi noterons-nous $\boldsymbol{\alpha}^{(t)}$, la concaténation des $\boldsymbol{\alpha}_k^{(t)}$, le vecteur[‡] de paramètres mis à jour à l'aide de $\mathcal{D}_t$.

[‡] Quoique la double indexation des α rappelle plus une matrice qu'un vecteur, la formulation d'un problème d'optimisation avec une matrice de paramètres n'est pas très courante

Ainsi, le théorème (9.1) nous assure qu'une minimisation de l'erreur régularisée définie par la somme de (9.6) et (9.14) par rapport à $f \in \mathcal{H}^N$ donnera lieu à un minimum *global* (notre fonction est convexe) qui pourra être paramétrisé par $\boldsymbol{\alpha}^{(t)}$.

Rappelons que cette formulation est extrêmement puissante, car le noyau nous permet ici de considérer un nombre immense, voire infini de caractéristiques en n'effectuant qu'un nombre fini d'opération. En regard de l'astuce du noyau (§ 7.4), nous pouvons considérer le noyau comme un paramètre libre dont nous tenterons d'optimiser l'usage.

9.3.1 Interprétation du théorème du représentant

Rappelons nous les équations (7.8) et (7.9)

$$\begin{aligned} \Phi : \mathcal{X} &\to \mathbb{R}^{\mathcal{X}} \\ x &\mapsto \mathrm{K}(x, .) \end{aligned}$$

En bref, le théorème du représentant nous affirme donc que si l'on se munit d'une base $\{\Phi(x_i)\}_{x_i \in \mathcal{X}_\mathcal{D}}$, alors les minima en f de $(\mathcal{L}^\Omega_\mathcal{D} f)$ seront dans l'espace de fonctions engendrées par ladite base. De même, le théorème nous apprend que le nombre de ces fonctions de base croît linéairement avec la taille de l'ensemble d'entraînement. L'utilisation d'une fenêtre (§ 8.1.1), limitant le nombre d'exemples aux Δ derniers, permet donc une forme de normalisation (Kivinen, Smola et Williamson 2002).

---

et, par conséquent, ni intuitive, ni très pratique. Bref, nous résumerons plutôt par

$$\boldsymbol{\alpha}_k^{(t)} = \left(\alpha_{k\tau}^{(t)}\right)_{\tau=t-(\Delta+h-1)}^{t-h} \in \mathbb{R}^\Delta$$

les paramètres associés au bien k au temps t, puis par

$$\boldsymbol{\alpha}^{(t)} = \left(\boldsymbol{\alpha}_k^{(t)}\right)_{k=1}^{N} \in \mathbb{R}^{N\Delta}$$

la concaténation des paramètres au temps t en un seul vecteur.

9.3.2 Optimisation sous-jacente au modèle

Nous reformulons finalement en termes de $\boldsymbol{\alpha}^{(t)}$ le problème d'optimisation qui sera au centre de l'entraînement de notre modèle

$$\min_{\boldsymbol{\alpha}^{(t)}} \; (\mathcal{L}_{\mathcal{D}_t} f)(\{x_{s-h}, y_s, f(x_{s-h})\}_{s=t-(\Delta+h-1)}^{t-h}) + \Omega(\, \|f\|_{pC}\,) \tag{9.31}$$

$$(\mathcal{L}_{\mathcal{D}_t} f)(\{x_{s-h}, y_s, f(x_{s-h})\}_{s=t-(\Delta+h-1)}^{t-h-1}) = -\overline{r}_P(\boldsymbol{\alpha}^{(t)}) + \frac{A}{2}\hat{\sigma}_P^2(\boldsymbol{\alpha}^{(t)}) \tag{9.32}$$

$$\overline{r}_P(\boldsymbol{\alpha}^{(t)}) = \frac{1}{\Delta} \sum_{s=t-(\Delta+h-1)}^{t-h} r_{P_s}(\boldsymbol{\alpha}^{(t)}) \tag{9.33}$$

$$\hat{\sigma}_P^2(\boldsymbol{\alpha}^{(t)}) = \frac{\Delta \sum_{s=t-(\Delta+h-1)}^{t-h} r_{P_s}(\boldsymbol{\alpha}^{(t)}) + \sum_{s=t-(\Delta+h-1)}^{t-h} r_{P_s}^2(\boldsymbol{\alpha}^{(t)})}{\Delta(\Delta-1)} \tag{9.34}$$

$$r_{P_s}(\boldsymbol{\alpha}^{(t)}) = \sum_{k=1}^{N} f_k(\boldsymbol{\alpha}^{(t)}, x_{s-h})\, r_{ki}(y_s) \tag{9.35}$$

$$\Omega(\, \|f\|_{pC}\,) = \gamma \sum_{k=0}^{N-1} \sum_{s=t-(\Delta+h-1)}^{t-h} (f_k(\boldsymbol{\alpha}^{(t)}, x_{s-h})\, p_{ks}(y_s))^2 \tag{9.36}$$

$$f_k(\boldsymbol{\alpha}^{(t)}, x_{s-h}) = \sum_{\tau=t-(\Delta+h-1)}^{t-h} \alpha_{k\tau}^{(t)}\, \mathrm{K}(x_{\tau-h}, x_{s-h}) \tag{9.37}$$

un problème quadratique sans contraintes — malgré les apparences. Ainsi, au temps t nous trouverons les paramètres globalement optimaux, ceux qui auraient permis d'obtenir, pour A et γ donnés, l'utilité maximale sur les périodes allant de $t-(\Delta+h-1)$ à $t-1$, compte tenu de la régularisation.

CHAPITRE 10

Cadre expérimental

Le chapitre précédent nous a permis d'introduire le modèle dans ses aspects théoriques. Dans le présent chapitre, nous établissons *les grandes lignes* des considérations pratiques rencontrées dans l'évaluation de la qualité d'un tel modèle. Tout au long du chapitre, nous tenterons de résumer le propos par des algorithmes de haut niveau.

10.1 Le défi

Comme ce modèle n'en est qu'à ses premières armes, nous avons cru bon de le tester sur le marché des contrats à terme boursiers, un marché où la dominance des investisseurs institutionnels limite en quelque sorte le bruit dans l'évolution des prix.

Dans le milieu financier, il est pratique courante d'évaluer les performances d'un gestionnaire de portefeuille à comparant le fruit de sa stratégie de placement avec la progression d'un indice de marché. Sur le marché qui nous intéresse, un indice reconnu est le *Mount Lucas Management Index* (indice MLM). Nous évaluerons donc notre modèle en comparant les performances

qu'il aurait obtenu sur les données historiques dont nous disposons aux performances de l'indice MLM sur les mêmes périodes.

Pour réaliser nos expériences, nous disposons d'une base de données contenant, pour les 25 contrats considérés par l'indice MLM (Tableau (10.1)), les informations quotidiennement issues du marché de contrats à terme boursiers du *Chicago Board of Trade* au cours des dernières années (3224 jours ouvrables ; 152 mois complets). À chaque jour, nous disposons, pour chacun des biens des informations suivantes : le prix d'ouverture, le plus haut et le plus bas prix de la journée, le prix de fermeture, le nombre de transactions dans la journée, l'indice de positions ouvertes, la date d'expiration du contrat, un indicateur quant à savoir si le bien est échangeable ladite journée et un autre quant à savoir s'il est temps de transférer vers une autre maturité de contrat.

Monnaies	**Énergie**	**Inst. Financiers**	**Grains**
Livre anglaise	Huile à chauffage	Bons du Trésor (US)	Maïs
Dollar Canadien	Essence sans plomb	Obligations 5 ans (US)	Fèves de soya
Euro	Pétrole Brut	Obligations 10 ans (US)	Mets à base de soya
Franc Suisse	Gaz Naturel		Huile de soya
Yen Japonais			Blé
Dollar Australien			
Métaux	**Exotiques**	**Viandes**	
Or	Café	Bétail	
Argent	Sucre		
Cuivre	Cotton		

Tableau 10.1 – Les Biens considérés par l'indice MLM

Nous considérerons le prix p_{kt} comme étant le prix à la fermeture de la journée t.

10.2 Entraînement sous le coût régularisé

Dans notre présentation de l'apprentissage supervisé (Chapitre 4), nous avons exposé les différents enjeux de l'entraînement d'un prédicteur, pour finalement conclure que la minimisation d'une erreur régularisée (Définition (4.4)) permettait d'obtenir un prédicteur adéquat en regard desdits enjeux. Conséquemment, nous avons établi au chapitre précédent un problème d'optimisation (Équations (9.31) à (9.37)) résumant l'erreur régularisée au cœur de notre modèle.

10.2.1 Signaux prédictifs et ACP

Tableau 10.2 – EXTRAIRE LES COMPOSANTES PRINCIPALES

Entrées: S ; n_{comp}

C	$\leftarrow$	Estimer à l'aide des données la matrice de covariance de la distribution sous-jacente aux points de S.
$(\boldsymbol{\lambda}, \mathrm{V})$	$\leftarrow$	Procéder à la décomposition en valeurs propres et vecteurs propres de la matrice C. Les éléments $[\boldsymbol{\lambda}]_l$ sont les valeurs propres de C ordonnées en ordre décroissant et les lignes de V, $[\mathrm{V}]_{l,:}$, sont les vecteurs propres correspondants au $[\boldsymbol{\lambda}]_l$.
$\tilde{\mathrm{S}}$	$\leftarrow$	Pour chaque $\mathbf{s} \in \mathrm{S}$ $\tilde{\mathbf{s}} \leftarrow$ Projeter $\mathbf{s}$ dans l'espace engendré par les n_{comp} premières lignes de V.

Sorties: $\tilde{\mathrm{S}}$

Dans le chapitre 8, nous avons présenté x_t comme étant l'information générée par le marché au temps t. Toutefois, il est fort peu probable que toutes ces informations soient prédictives à l'état brut. Comment distinguer l'information qui est prédictive de celle qui ne l'est pas ? Loin de vouloir nous attaquer au défi de répondre à cette question, nous nous en remettons plutôt aux réponses apportées par moult travaux issus de la communauté économique et d'ap-

prentissage automatique (CHAPADOS 2000). Entre autres signaux d'intérêt, les rendements et log-rendements reviennent fréquemment ; le MLM, que nous tentons de concurrencer, est aussi bâti sur des signaux indiquant la tendance des rendements.

Quoi qu'il en soit, si pour chaque bien nous considérons n_{sign} signaux relatifs aux rendements des différents biens, nous aurons $x_t \in \mathbb{R}^n$, avec $n = n_{\text{sign}}N$. Pour $n_{\text{sign}} = 6$ un nombre somme toute raisonnable, on se retrouve dans un espace de départ $\mathcal{X}$ de dimension n =300 ($N = 25$). En regard des arguments présentés au chapitre 6, il y a fort à parier que notre puissance de généralisation serait relativement limitée. De plus, les méthodes à noyaux sont toutes relativement coûteuses en temps de calcul (habituellement dans $O(n^3)$). Il est donc déraisonable d'envisager travailler directement avec $x_t \in \mathbb{R}^{300}$.

Pour réduire la dimension de $\mathcal{X}$ nous appliquerons donc une ACP (§ 5.2) aux signaux présélectionnés. La possibilité de surapprentissage* et le temps de calcul s'en trouveront ainsi réduits. Remarquons que cette procédure entraîne l'ajout d'un nouvel hyperparamètre†, n_{comp}, qu'il faudra choisir avec soin (§ 10.4).

10.3 Évaluation de la performance par validation séquentielle

Afin d'évaluer notre modèle, la solution la plus intuitive est de se doter d'une mesure objective de la performance qu'aurait historiquement obtenu le modèle

* En éliminant les dimensions les moins représentatives, l'ACP permet de réduire l'apprentissage du bruit dans les données.

† En apprentissage, on appelle **hyperparamètre** un paramètre qui n'est pas appris par le modèle, mais qui doit être choisi par ailleurs. Jusqu'à maintenant, K, le noyau ; A, l'aversion au risque ; γ, le paramètre de régularisation ; Δ, la fenêtre d'entraînement ; et h, l'horizon ; sont de tels paramètres non appris mais néanmoins indispensables à la définition d'un prédicteur sous notre modèle.

Tableau 10.3 – ENTRAÎNER LE PRÉDICTEUR $\mathcal{A}$

Entrées: t ; $\mathcal{D}_t$		
$\tilde{\mathcal{X}}_{\mathcal{D}_t}$	$\leftarrow$	EXTRAIRE LES COMPOSANTES PRINCIPALES($\mathcal{X}_{\mathcal{D}_t}$, n_{comp}).
$\tilde{\mathcal{D}}_t$	$\leftarrow$	$\{(\tilde{x}_{s-h}, y_s)\}_{s=t-(\Delta+h-1)}^{t-h}$
$\boldsymbol{\alpha}^{(t)}$	$\leftarrow$	Utiliser le gradient conjugué (CONCUS, GOLUB et O'LEARY 1976) pour annuler le gradient de $(\mathcal{L}_{\tilde{\mathcal{D}}_t}^{\Omega} f)$ (Équation (9.31)).
Sorties: $\boldsymbol{\alpha}^{(t)}$		

sur les marchés financiers. Pour ce faire, il faut d'abord et avant tout se doter d'une interface avec la réalité des marchés permettant de juger sans biais de ladite performance.

10.3.1 Simulation d'un marché

Bien que nous considérions un modèle simple, ne recommendant qu'un portefeuille idéal, il est primordial d'intégrer les frictions du marché à l'évaluation des performances. Ainsi, au moment de calculer les rendements, les coûts de transaction pour passer d'un portefeuille à l'autre, l'impossibilité de transiger d'un bien donné certaines journées, etc, doivent être considérés.

Pour ce faire, nous avons dû implémenter un module d'échanges qui simule la réalité du marché. Dans cette réalité, il est important d'avoir conscience que le test comporte deux parties chronologiquement distinctes. D'abord, le prédicteur considère l'observation x_{t-h}, information disponible au temps $(t-h)$ et émet une prédiction, dans notre modèle $\mathbf{w}_t$. Ensuite, il est possible d'utiliser y_t, disponible au temps t, pour évaluer la qualité de la prédiction à $(t-h)$, $\mathbf{w}_t$.

Tableau 10.4 – TESTER LE PRÉDICTEUR $\mathcal{A}$

Entrées: t ; $\boldsymbol{\alpha}^{(t)}$; (x_{t-h}, y_t)		
$\mathbf{w}_t$	$\leftarrow$	$f(x_{t-h}; \boldsymbol{\alpha}^{(t)})$: la prédiction, qui ne tient pas compte de y_t.
r_{P_t}	$\leftarrow$	Évaluer le rendement du portefeuille $\mathbf{w}_t$ sur la période t. Soustraire du rendement les frais de transaction encourus pour passer de $\mathbf{w}_{t-1}$ à $\mathbf{w}_t$.
Sorties: r_{P_t}		

10.3.2 Validation séquentielle

Lorsque l'on travaille avec des séries chronologiques, non seulement l'ordre dans lequel on considère les données nous est-il fondamentalement imposé, mais la distribution sous-jacente aux données n'est pas nécessairement constante à travers le temps. Cette section a donc pour but de présenter un estimateur de l'erreur de généralisation par **validation séquentielle**, introduit[‡] par (GINGRAS, BENGIO et NADEAU 1999), adapté à l'analyse de telles séries temporelles.

Tableau 10.5 – VALIDATION SÉQUENTIELLE (Figure (10.1))

Entrées: $\{x_t\}_{t=1}^{T}$; $\{y_t\}_{t=1}^{T}$; Un prédicteur $\mathcal{A}$; L'indice $t_0 > \Delta + h - 1$ du premier temps de test.		
$\{r_{P_t}\}_{t=t_0}^{T-h}$	$\leftarrow$	Pour t allant de t_0 à $T-h$ (incréments de 1) $\mathcal{D}_t \leftarrow \{(x_{\tau-h}, y_\tau)\}_{\tau=t-(\Delta+h-1)}^{t-h}$ $\boldsymbol{\alpha}^{(t)} \leftarrow \mathcal{A}.\text{ENTRAÎNER}(t, \mathcal{D}_t)$ $r_{P_t} \leftarrow \mathcal{A}.\text{TESTER}(t, \boldsymbol{\alpha}^{(t)}, (x_{t-h}, y_t))$
Sorties: $\{r_{P_t}\}_{t=t_0}^{T-h}$		

[‡]Sous le nom original de validation croisée séquentielle (*Sequential Cross-Validation*)

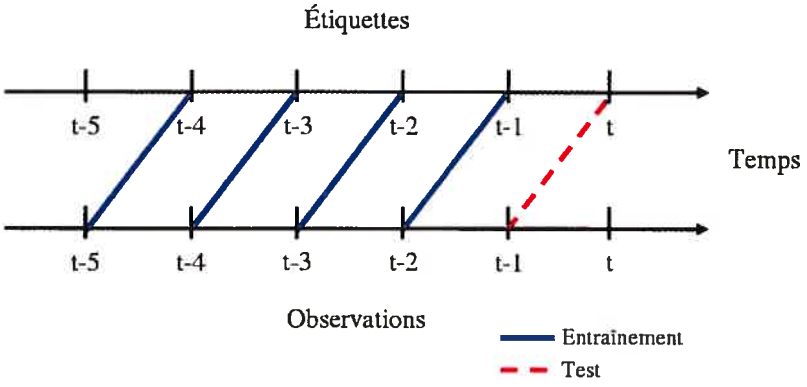

Figure 10.1 – *À l'itération t, l'ensemble d'entraînement est constitué des Δ paires attachées ici par une double ligne. Par la suite, le prédicteur est tester sur la paire suivante, attachée ici en pointillés. Sur ce schéma, $h = 1$ et $\Delta = 4$.*

Cet algorithme (Tableau (10.5)) mise sur le fait que $f(.; \boldsymbol{\alpha}^{(t)})$ sont près les unes des autres au sens où les paramètres $\boldsymbol{\alpha}^{(t)}$ ne devrait pas changer drastiquement d'une période à l'autre, surtout pour h petit (CHAPADOS et BENGIO 2003).

D'autre part, même s'il est fort probable que cet estimateur en soit un biaisé de l'erreur de généralisation, il n'en demeure pas moins qu'il évalue la perte réellement encourue par l'algorithme. Dans un contexte financier, ce réalisme n'est pas pour déplaire.

10.4 Sélection des hyperparamètres

Évidemment, pour chaque ensemble d'hyperparamètres $\{\mathrm{K}, \mathrm{A}, \gamma, \Delta, h, n_{\text{comp}}\}$, notre modèle, $\mathcal{M}$, donne lieu à un prédicteur, $\mathcal{A}$, différent :

$$\mathcal{A} = \mathcal{M}(\{\mathrm{K}, \mathrm{A}, \gamma, \Delta, h, n_{\text{comp}}\}) \tag{10.1}$$

où l'on considère le modèle comme une fonction qui, étant donné un ensemble d'hyperparamètres, retourne un prédicteur défini à l'aide de ses derniers.

Par ailleurs, l'algorithme de validation séquentielle (Tableau (10.5)), accepte en entrée un prédicteur $\mathcal{A}$ bien défini afin d'en évaluer la qualité. Il va sans dire, un choix judicieux de ses hyperparamètres doit être fait au préalable.

Dans le cadre de ce mémoire, nous utiliserons pour ce faire un hybride de validation croisée (§ 4.4) et de validation séquentielle, une méthodologie éprouvée (CHAPADOS et BENGIO 2003). Nous allons donc séparer la base de donnée à laquelle nous avons accès en trois ensembles d'entraînement, de validation et de test. Pour la sélection de paramètres, nous utiliserons donc les ensembles

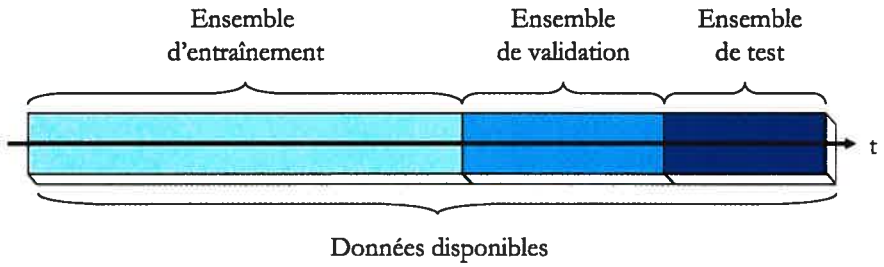

Figure 10.2 – *Séparation de la bases de données en ensembles d'entraînement, de validation et de test.*

d'entraînement et de validation, afin de pouvoir faire un test final non biaisé sur l'ensemble de test.

Avant de présenter l'algorithme (Tableau (10.6)), remarquons que nous utiliserons pour mesure de performance le ratio de Sharpe défini tout simplement par $\mathrm{SR}(\{r_{P_t}\}) = \frac{\bar{r}(\{r_{P_t}\})}{\hat{\sigma}(\{r_{P_t}\})}$. Ce dernier est effectivement proportionnel au rendement et inversement proportionnel au risque inhérent à notre stratégie. Nous ne pouvions évidemment pas considérer l'utilité comme critère de performance puisqu'elle dépend de l'aversion au risque A, un hyperparamètre du modèle.

Après avoir estimé les hyperparamètres optimaux $\{\mathrm{K}, \mathrm{A}, \gamma, \Delta, h, n_{\mathrm{comp}}\}^*$ à l'aide de l'algorithme de de sélection des hyperparamètres (Tableau (10.6)), nous

Tableau 10.6 – SÉLECTION DES HYPERPARAMÈTRES

Entrées: Un modèle $\mathcal{M}$; $\{x_t\}_{t=1}^{\mathbf{T}}$; $\{y_t\}_{t=1}^{\mathbf{T}}$; ainsi que $\mathbf{K}$, $\mathbf{A}$, Γ, $\mathbf{\Delta}$, $\mathbf{h}$ et $\mathbf{n_{comp}}$ des ensembles de valeurs admissibles pour K, A, γ, Δ, h et n_{comp} respectivement.

T_{ent} $\leftarrow$ Choisir la proportion p de l'ensemble à conserver exclusivement pour l'entraînement et trouver $T_{\text{ent}} < \mathbf{T}$ tel que $\frac{T_{\text{ent}}}{\mathbf{T}} \approx p$.

T_{valid} $\leftarrow$ Choisir la proportion p de l'ensemble à utiliser pour valider le choix des hyperparamètres et trouver $T_{\text{valid}} < \mathbf{T}$ tel que $\frac{T_{\text{valid}} - T_{\text{ent}}}{\mathbf{T}} \approx p$.

S* $\leftarrow$ Pour tout S $\in \mathbf{K} \times \mathbf{A} \times \Gamma \times \mathbf{\Delta} \times \mathbf{h} \times \mathbf{n_{comp}}$

$\mathcal{A} \leftarrow \mathcal{M}(\text{S})$

$\{r_{P_t}\}_{t=T_{\text{ent}}}^{T_{\text{valid}}-h} \leftarrow$ VALIDATION SÉQUENTIELLE$(\{x_t\}_{t=1}^{T_{\text{valid}}}, \{y_t\}_{t=1}^{T_{\text{valid}}}, \mathcal{A}, T_{\text{ent}})$

Évaluer le ratio de Sharpe sur la série de rendements obtenus et comparer à ceux obtenus précedemment : conserver dans S* l'ensemble ayant obtenu le meilleur ratio de Sharpe jusqu'à présent.

Sorties: T_{valid} ; $\{\text{K}, \text{A}, \gamma, \Delta, h, n_{\text{comp}}\}^*$

pourrons finalement utiliser le prédicteur

$$\mathcal{A}^* = \mathcal{M}\left(\{\text{K}, \text{A}, \gamma, \Delta, h, n_{\text{comp}}\}^*\right) \tag{10.2}$$

et obtenir la série des rendements de test

$$\{r_{P_t}\}_{t=T_{\text{valid}}}^{\mathbf{T}-h} \leftarrow \text{VALIDATION SÉQUENTIELLE}(\{x_t\}_{t=1}^{\mathbf{T}}, \{y_t\}_{t=1}^{\mathbf{T}}, \mathcal{A}^*, T_{\text{valid}}) \tag{10.3}$$

pour enfin pouvoir évaluer la qualité de notre modèle en comparaison à l'indice MLM.

CHAPITRE 11

Résultats

Le présent chapitre introduit les résultats préliminaires obtenus sous le cadre expérimental établi au chapitre précédent. Les résultats seront analysés et, une fois les hyperparamètres choisis, la performance de test sera comparée à celle de l'indice MLM sur la même fenêtre de temps.

11.1 Sans frais de transaction

Nous considérerons la période s'écoulant entre les temps $(t-1)$ et t comme étant un mois. Les résultats présentés dans cette section *ne tiennent pas compte* des frais de transaction.

11.1.1 Sélection des hyperparamètres

Tout au long de ce chapitre, nous nous concentrerons sur $h = 1$, puisque nous disposons de toutes les données nécessaires à chaque période pour évaluer la qualité de la précédente prédiction. Nous nommerons ensemble d'entraînement les 74 premiers mois disponibles dans la base de données et ensemble de valida-

tion les mois 75 à 104 (30 mois). Les derniers 48 mois constituerons l'ensemble de test.

Rappelons nous que nous devons choisir, à l'aide des ensembles d'entraînement et de validation,

$$\{\mathrm{K}, \mathrm{A}, \gamma, \Delta, n_{\mathrm{comp}}\} \in \mathbf{K} \times \mathbf{A} \times \Gamma \times \mathbf{\Delta} \times \mathbf{n_{comp}} \tag{11.1}$$

où K est le noyau sous-jacent au prédicteur, A, l'aversion au risque, γ, le paramètre de régularisation et n_{comp}, le nombre de composantes à conserver lors de l'ACP sur les signaux de rendement. Clarifions ici que nous utiliserons, à chaque temps t et pour chaque bien k, des signaux de rendement définis de la manière suivante :

$$\tilde{x}_t = (\mathrm{sign}(\zeta_{kt,l}), \tanh(a\,\zeta_{kt,l}))_{(a,l)\in\{0.5,1,2,5,10\}\times\{6,12,24\}} \tag{11.2}$$

où

$$\zeta_{kt,l} = \frac{p_{kt} - \nu_{kt,l}}{\nu_{kt,l}} \tag{11.3}$$

$$\nu_{kt,l} = \frac{1}{l}\sum_{s=t-l+1}^{t} p_{ks}. \tag{11.4}$$

Les observations $x_t \in \mathbb{R}^{n_{\mathrm{comp}}}$ fournies au noyau seront donc les projections des $\tilde{x}_t \in \mathbb{R}^{450}$ dans l'espace engendré par leur n_{comp} vecteurs propres principaux.

Nombre de composantes principales dans l'ACP

Devant l'immensité de l'espace à explorer, il n'est pas nécessairement évident de savoir par où commencer. Quoiqu'il en soit, nous pouvons débuter par analyser le comportement des composantes principales à travers le spectre de la matrice de covariance empirique des signaux de rendements.

Suite à l'observation de la figure (11.1), nous pouvons déjà restraindre $\mathbf{n_{comp}}$ à l'ensemble $\{5, \ldots, 25\}$. En effet, après la 25^e valeur propre, l'importance

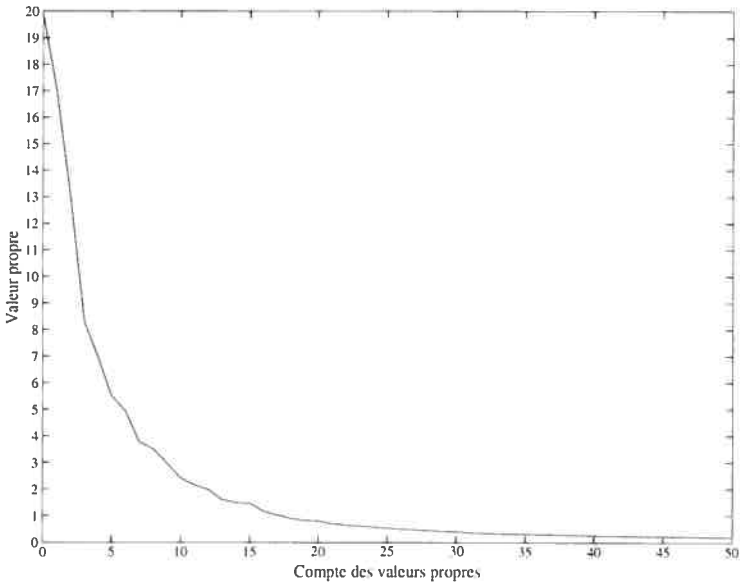

Figure 11.1 – *Analyse du spectre de la matrice de covariance empirique des signaux de rendements. Nous considérons 450 signaux sur les rendements. Cependant, l'importance relative de chaque valeur propre diminue rapidement.*

relative de chaque nouvelle dimension est presque nulle et, conséquemment, l'information ajoutée n'est probablement plus pertinente.

Choix du noyau

Cette première restriction aidant, il demeure la question du choix du noyau. Comme son utilisation est très répandue dans les applications financières, nous utiliserons un noyau gaussien

$$\mathrm{K}(x, y) = e^{\frac{-\|x-y\|^2}{\sigma^2}} \tag{11.5}$$

Reste qu'en choisissant ce noyau nous devons maintenant prendre soin de choisir $\sigma^2 \in \mathbb{R}_+$, car le noyau gaussien est très sensible à la valeur de σ. Si σ est

A	γ	σ	Entraînement	Validation
1.75	0.0007	2.5	1.020	0.707
1.75	0.001	2	0.654	0.635
1.00	0.001	2	0.594	0.631
1.50	0.001	2.5	0.882	0.627
0.75	0.0007	2	0.599	0.625
1.25	0.001	2.5	0.834	0.623
1.25	0.0007	2	0.659	0.623
1.50	0.001	2	0.635	0.622
1.00	0.0007	2	0.629	0.621
1.50	0.0007	2	0.672	0.621
1.75	0.0007	2	0.679	0.620
1.75	0.001	2.5	0.914	0.618
1.00	0.0007	2.5	0.864	0.618
1.25	0.001	2	0.617	0.618
1.25	0.0007	2.5	0.924	0.616
0.75	0.001	2	0.560	0.615
0.75	0.0007	2.5	0.801	0.612
1.00	0.001	2.5	0.792	0.605
0.75	0.001	2.5	0.744	0.591
0.75	0.001	3	0.706	0.538
1.00	0.0007	3	0.859	0.501
1.25	0.0007	3	0.904	0.494
1.00	0.001	3	0.763	0.474
1.00	0.0007	1	0.021	0.472
1.50	0.001	3	0.875	0.471
1.75	0.0007	3	1.026	0.467
0.75	0.0007	3	0.781	0.464
1.25	0.0007	2.5	0.031	0.447
1.00	0.0007	2	0.021	0.444
1.50	0.0007	3	0.969	0.437

Tableau 11.1 – Des 270 prédicteurs issus du treillis original, nous reportons les performances en entraînement et en validation des 30 meilleurs (validation). Lesdits résultats *ne tiennent pas compte* des frais de transaction.

trop petit, $\mathrm{K}(x, y)$ deviendra rapidement 1 et notre prédicteur sera enclin au surapprentissage. Par contre, pour σ trop grand, $\mathrm{K}(x, y)$ devient rapidement 1 et, ce faisant, le prédicteur n'a pas plus de puissance qu'un banal prédicteur linéaire.

Exploration de l'espace

Pour aller plus d'avant dans notre sélection de paramètres, nous irons d'une première exploration gloutone de l'espace. En utilisant $n_{\text{comp}} = 10$, $\Delta = 60$ mois (5 ans ; une fenêtre souvent utilisée par les économistes) et en limitant Γ à $\{0.001, 0.0007\}$*, Σ à $\{1.0, 1.5, \ldots, 5\}$ et $\mathbf{A}$ à $\{0.5, 0.75, 1.0, \ldots, 4\}$ nous espérons couvrir l'espace vraisemblable des solutions d'un premier treillis, que nous raffinerons petit à petit.

Déjà, les résultats présentés dans le tableau (11.1) nous laissent croire que σ devrait être environ 2 ou 2.5. Conséquement, nous avons décidé, en fixant $(\mathrm{A}, \gamma) = (1.75, 0.0007)$, d'étudier l'influence commune du nombre de composantes principales et de σ.

Les résultats de 94 expériences sur l'interaction de n_{comp} et σ sont présentés à la figure (11.3). Sur le premier graphe, nous pouvons observer que la performance de validation semble être au mieux avec 10 ou 11 composantes principales et une valeur de 2.5 pour σ. En effet, pour ces valeurs de n_{comp}, la performance semble croître avec σ sur l'intervalle 1.75 à 2.5 pour ensuite chuter avec $\sigma = 2.75$. Le second graphe, plus raffiné, permet de voir que le nombre de composantes optimal est autour de 10. La performance augmente de $\sigma = 2.3$ à $\sigma = 2.4$ pour ensuite redescendre. Conséquemment, nous opterons pour $(n_{\text{comp}}, \sigma) = (10, 2.4)$ pour la suite des expériences.

*Les prix moyens des contrats sont de l'ordre des miliers de dollars ; une fois mis au carrés, on atteint l'ordre du million. Comme nous espérons des rendements moyens de l'ordre des miliers de dollars, il est raisonable de croire qu'un facteur de régularisation de l'ordre du millième permettra de garder la pénalisation dans le même ordre de grandeur que les revenus. Nous laissons à plus tard une étude exhaustive l'influence du paramètre de régularisation.

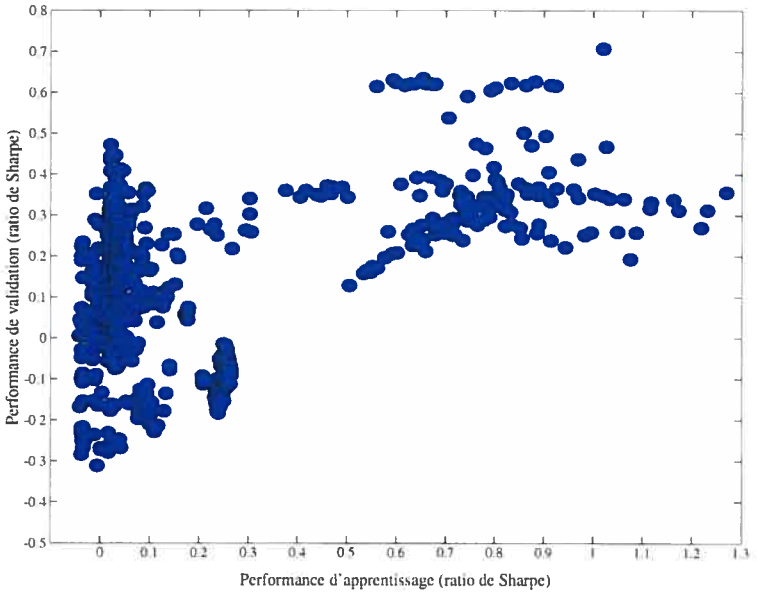

Figure 11.2 – *Performance de validation en fonction de la performance d'entraînement sur 270 expériences : Comme nous en avions discuté dans les sections 4.3 et 4.4, l'erreur d'apprentissage n'est qu'un estimateur bruité de la performance de généralisation.*

Sur ce graphique, nous pouvons constater que la performance de validation est loin d'être monotone suivant l'augmentation de la performance d'entraînement. Quoiqu'il en soit, une certaine tendance à la hausse est tangible jusque aux environ de 1 où le modèle semble commencer à surapprendre.

Remarquons aussi que la performance de validation de 0.707 (première dans le tableau (11.1)) semble être une observation aberrante. Cette performance est peut-être due au hasard plus qu'à la qualité des paramètres choisis.

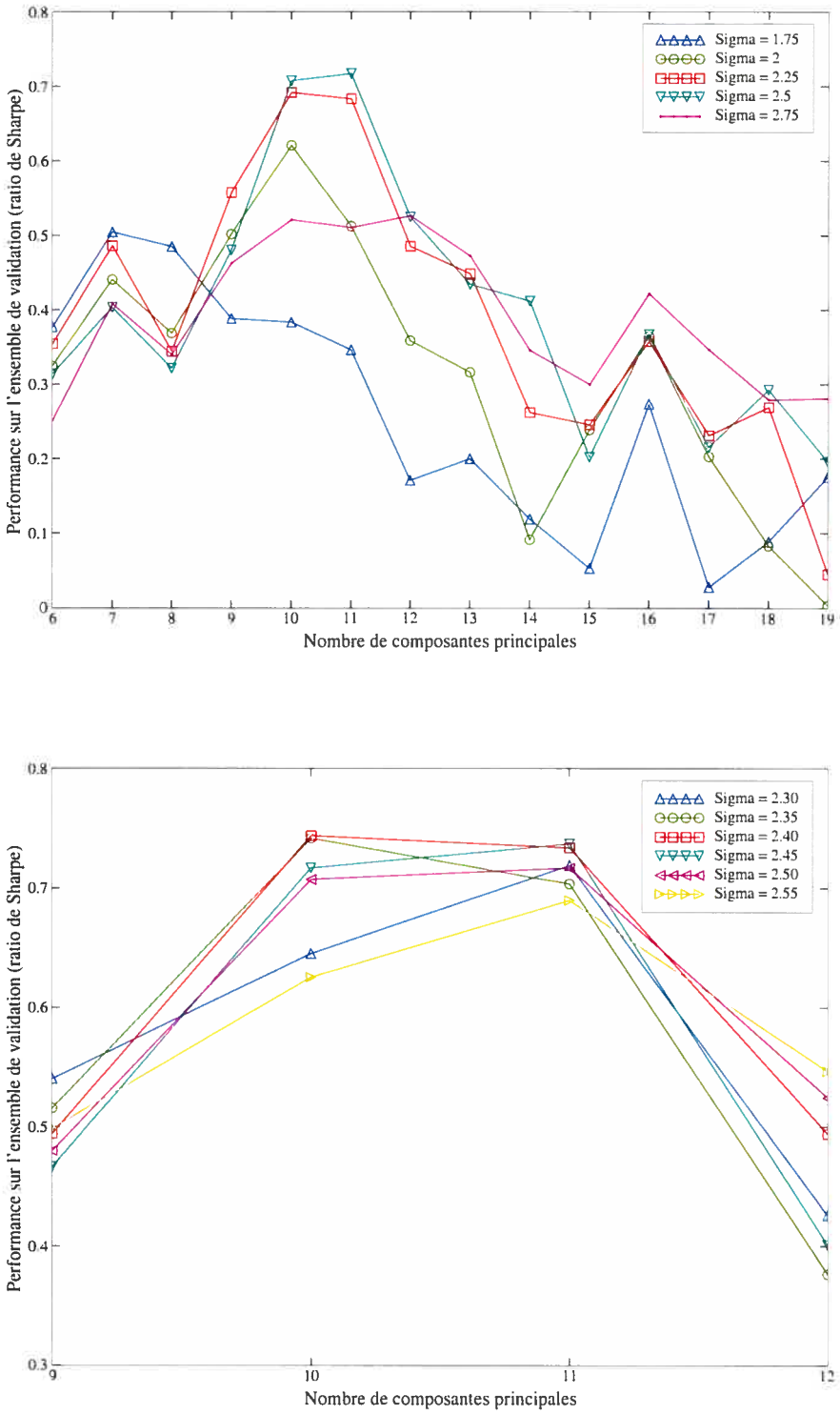

Figure 11.3 – *Influence conjointe de n_{comp}et de σ*

Aversion pour le risque et paramètre de régularisation

Encore une fois, nous analyserons conjointement l'influence de deux hyper-paramètres, l'aversion pour le risque A et le paramètre de régularisation γ. Cette analyse conjointe est d'autant plus justifiée qu'une variation de l'un de ces paramètres influence l'importance relative du second dans la fonction de coût.

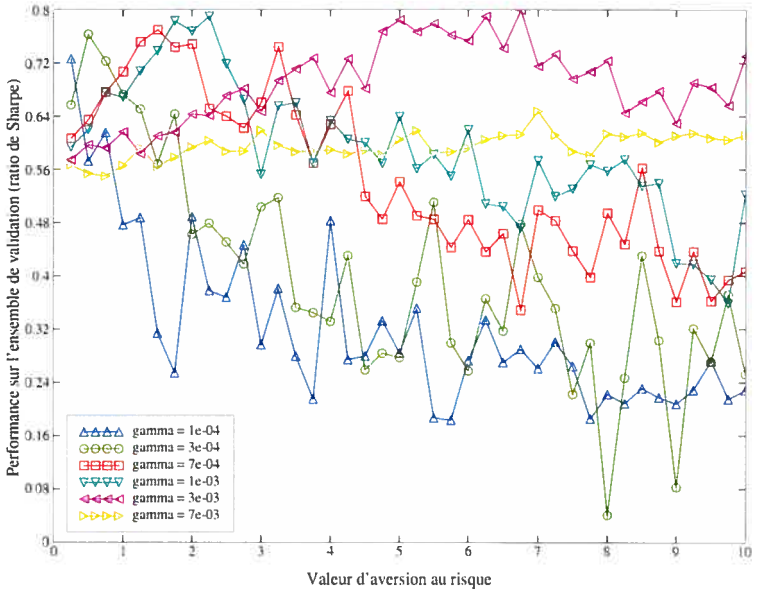

Figure 11.4 – *Influence conjointe de l'aversion au risque et du paramètre de régularisation : Ratio de Sharpe*

Au premier regard, la figure (11.4) ne semble pas très instructive, sauf pour montrer que $(A, \gamma) = (2.25, 0.001)$ ou $(A, \gamma) = (6.75, 0.003)$ obtiennent les meilleurs ratios de Sharpe en validation. Toutefois, rappelons-nous la figure (11.2). Il est possible qu'une performance de 0.8 soit tout simplement due à quelques "bon coups" d'un prédicteur sur un ensemble de validation qui lui est favorable. Conséquemment, un ratio de Sharpe près de 0.8 en validation n'est pas nécessairement garant de la capacité de généralisation du prédicteur. De

surcroît, d'un point de vue strictement économique, un ratio de Sharpe de 0.8 est pratiquement une aberration. De la figure (11.4), nous ne tirons donc pas de conclusion hâtive ; nous nous en inspirons simplement pour avoir une idée des combinaisons (A, γ) qui peuvent atteindre une performance de validation d'environ 0.6, plus raisonable en regard de la figure (11.2).

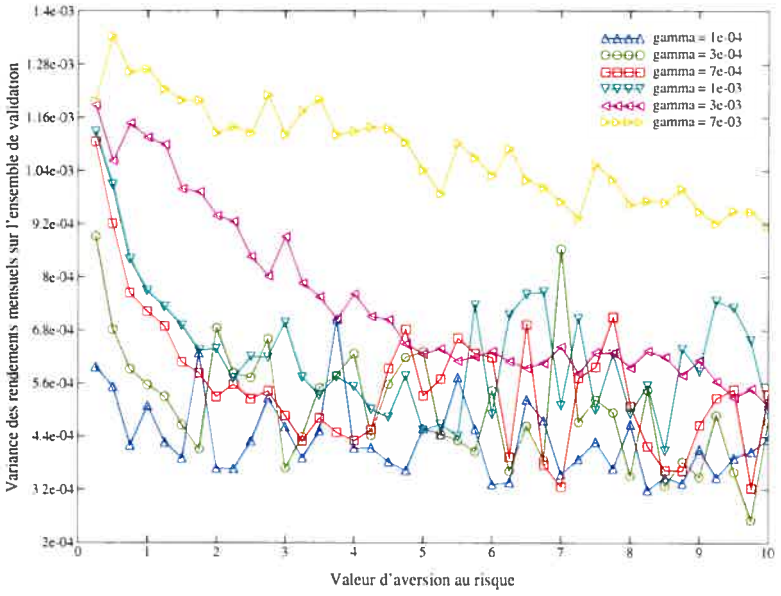

Figure 11.5 – *Influence conjointe de l'aversion au risque et du paramètre de régularisation : Variance des rendements.*

L'observation d'un ratio élevé peut évidemment avoir deux causes : un rendement élevé ou une variance faible. Des deux scénarios, en autant que les rendements demeurent attirants, un investisseur préférera souvent diminuer la variance, amenuisant ainsi le risque qu'il court. Pour discriminer parmis les combinaisons (A, γ) qui obtiennent un ratio de Sharpe satisfaisant (Figure (11.4)) nous tenterons de voir lesquelles le font via une faible variance (Figure (11.5)). À l'étude conjointe des deux graphes, nous remarquons qu'en utilisant une aversion pour le risque de 4 et un paramètre de régularisation de 7×10^{-4}, nous obtenons une performance de validation plus que respectable (ratio de

sharpe de 0.628 ; rendement mensuel moyen annualisé 16.8%) tout observant l'une des variances les moins élevées (4.31×10^{-4}, mensuellement). Dès lors, nous utiliserons les précédentes valeurs au cours des expériences à venir.

Validation de la fenêtre

Au début de notre sélection des hyperparamètres, nous avons choisi de fixer notre fenêtre d'entraînement à $\Delta = 60$ mois question de se concentrer sur le choix des autres hyperparamètres. Nous avions alors justifié notre choix en remarquant que moult méthodes pratiques utilisées par les économistes utilisaient une fenêtre de 5 ans. Pour s'assurer de la validité de ce choix,

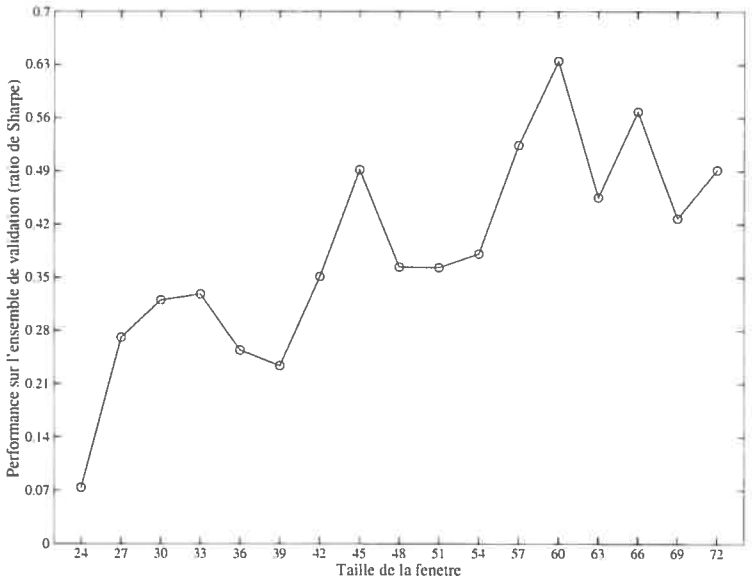

Figure 11.6 – *Performance de validation en fonction de la taille de la fenêtre. Nous avons utilisé 10 composantes principales, une aversion au risque* $A = 4$, *un paramètre de régularisation* $\gamma = 7 \times 10^{-4}$ *et un noyau gaussien avec* $\sigma = 2.4$.

nous avons, comme pour les autres hyperparamètres, isolé l'influence de Δ en effectuant une série d'expériences où les autres paramètres étaient fixes

par ailleurs (10 composantes principales, une aversion au risque A = 4, un paramètre de régularisation $\gamma = 7 \times 10^{-4}$ et un noyau gaussien avec $\sigma = 2.4$).

Il n'est pas surprenant de voir $\Delta = 60$ surpasser en performance (Figure (11.6)) les autres tailles de fenêtres. Rappelons-nous que le marché des contrats à terme boursiers en est un dominé par les investisseurs institutionnels. La plupart d'entres eux utilisent des méthodes quantitatives. En se fiant sur les études pratiques diffusées par les économistes, les investisseurs en présence finissent plus souvent qu'autrement par opter pour une fenêtre de 5 ans. Conséquemment, si cette fenêtre n'était pas intrinsèque à l'évolution du marché initialement, elle a fini par le devenir.

11.1.2 Performance de test

Nous somme désormais convaincus que l'utilisation d'une fenêtre de 60 mois, de 10 composantes principales, d'une aversion au risque de A = 4, d'un paramètre de régularisation $\gamma = 7 \times 10^{-4}$ et d'un noyau gaussien avec $\sigma = 2.4$ mène à un prédicteur adapté à la réalité du marché que nous étudions. Ainsi, nous sommes à même de comparer, sur l'ensemble de test de 48 mois, les performances de notre modèle relativement à l'indice de marché MLM.

Un premier regard au tableau (11.2) permet de constater que notre modèle obtient, sur l'ensemble de test, un meilleur ratio de Sharpe que l'indice MLM. Cependant, on ne peut se fier aveuglément à ce ratio en tant que mesure de performance. En effet, le ratio de Sharpe est utile pour comparer entre elles deux stratégies de placement au rendements similaires, mais il faut être conscient que cette mesure a ses limites. De fait, un investissement au taux sans risque présente, théoriquement, un ratio de Sharpe infini, argument qui n'aidera toutefois personne à convaincre les investisseurs de se retirer des marchés risqués...

Il est donc important de comparer entre elles les suites de rendements des deux modèles. Voilà pourquoi, dans le tableau (11.2), nous avons calculé la t-statistique issue d'un test pairé sous l'hypothèse nulle que les rendements de

	Modèle	Moyenne	Variance	Ratio de Sharpe
Complets t= 2.296	Indice MLM	3.70×10^{-3}	1.17×10^{-4}	0.342
	Notre modèle	1.47×10^{-2}	1.38×10^{-3}	0.397
Partiels t= 2.972 (sans 37)	Indice MLM	4.18×10^{-3}	1.09×10^{-4}	0.401
	Notre modèle	1.73×10^{-2}	1.09×10^{-3}	0.525

Tableau 11.2 – Résultats de l'indice MLM et de notre prédicteur sur l'ensemble de test. La valeur t est en fait la t-statistique résultant d'un test pairé sur les rendements. L'hypothèse nulle est que notre modèle ne fait pas mieux que le MLM, laquelle hypothèse peut être rejetée avec un niveau de confiance de 97.5%. Les résultats dits *partiels* sont ceux que l'on aurait obtenus en retirant l'observation 37.

notre modèle n'étaient pas supérieurs à ceux de l'indice. Avec une valeur de 2.296, nous pouvons affirmer que nos rendements sont statistiquement significativement supérieurs à ceux de l'indice à un niveau de confiance de 97.5%.

La figure (11.7) présente les rendements obtenus sur les 48 mois de test. Évidemment, les rendemments de notre modèle dominent ceux de l'indice, mais nous observons aussi un lacune du modèle sur ce graphique. Au 37^e mois de test, le modèle essuit une perte aberrante.

Cette perte se produit simultanément à une baisse de l'indice, lequel établit ses positions en suivant la tendance du marché. Il est donc possible, qu'au 37^e mois de test, le marché ait, par exemple, effectué un recul inattendu dû à l'explosion d'une bulle spéculative, à la rarification soudaine de l'un des sous-jacents, ... Peu importe la cause, nous constatons que notre modèle est fortement vulnérable à ce genre de renversements. Il serait probablement bénéfique que notre modèle soit jumelé à de quelconques règles de filtres ou intégré à un comité d'experts (CHAPADOS 2000). En effet, nous avons reporté dans le tableau (11.2) (résultats *partiels*), la performance qu'aurait connu notre modèle

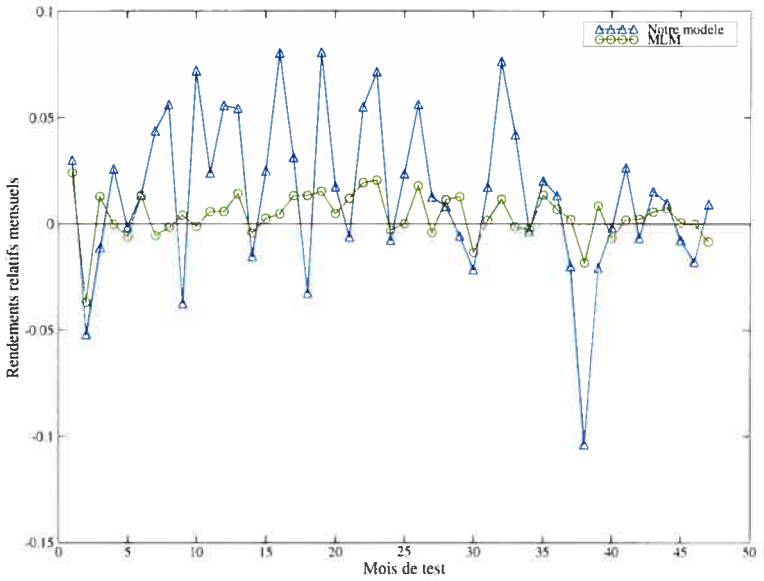

Figure 11.7 – *Rendements comparés de l'indice MLM et de notre modèle sur l'ensemble de test.*

si une quelconque source extérieure lui avait évité ce faux pas. Il est clair que cette contre-performance ponctuelle est très coûteuse au modèle.

11.2 Avec frais de transaction

Révisons maintenant la performance, *avec frais de transaction*, du prédicteur utilisant un noyau gaussien ($\sigma = 2.4$), une fenêtre Δ de 60 mois, les 10 composantes principales des signaux de rendements et une utilité régularisée où $A = 4$ et $\gamma = 7 \times 10^{-4}$. Le tableau (11.3) et la figure (11.8) présente les résultats obtenus avec des frais de transaction multiplicatif[†].

[†] Pour que son courtier effectue la transaction, l'investisseur doit débourser, pour chaque contrat acheté ou vendu, 2.80$ (s'il s'agit d'un contrat transigé électroniquement) ou 6.99$

	Modèle	Moyenne	Variance	Ratio de Sharpe
Complets	Indice MLM	3.70×10^{-3}	1.17×10^{-4}	0.291
t= 2.33				
	Notre modèle	1.47×10^{-2}	1.38×10^{-3}	0.387
Partiels	Indice MLM	4.18×10^{-3}	1.09×10^{-4}	0.346
t= 3.02				
(sans 37)	Notre modèle	1.73×10^{-2}	1.09×10^{-3}	0.514

Tableau 11.3 – Résultats de l'indice MLM et de notre prédicteur sur l'ensemble de test, avec frais de transaction. La valeur t est la t-statistique résultant d'un test pairé sur les rendements. L'hypothèse nulle est que notre modèle ne fait pas mieux que le MLM, laquelle hypothèse peut être rejetée avec un niveau de confiance de 99%. Les résultats dits *partiels* sont ceux que l'on aurait obtenus en retirant l'observation 37.

À prime abord, on pourrait argumenter que le précédent prédicteur à été sélectionné dans un cadre où nous ne tenions pas compte des frais de transaction et qu'une nouvelle sélection s'impose. Loin d'être saugrenu, cet argument est tout à fait sensé et serait effectivement d'application si les frais de transaction étaient plus imposants. Or, dans le marché des contrats à terme boursiers, les frais de transactions sont très modestes et n'aurait eu, par conséquent, que très peu d'influence sur le choix du prédicteur.

Dans le même ordre d'idée, vue les faibles frais de transaction, notre modèle performe très bien malgré son amnésie en regard des portefeuilles précédents, même que sa domination sur l'indice MLM s'accentue (Tableau (11.3)).

(s'il s'agit d'un contrat transigé sur le parquet). Dépendemment des contrats, ces montants représente entre 0.005% et 0.1% de la valeur du contrat.

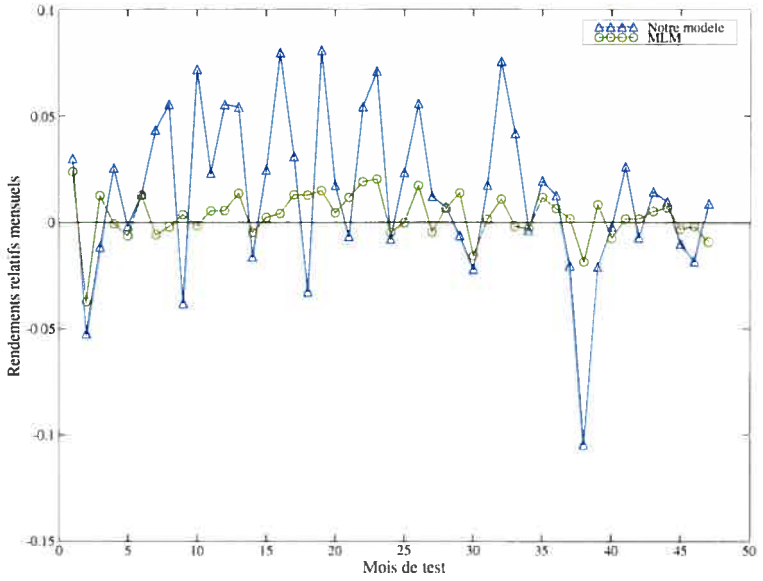

Figure 11.8 – *Rendements comparés de l'indice MLM et de notre modèle sur l'ensemble de test, avec frais de transaction.*

11.2.1 Portefeuille optimal sans égard au portefeuille précédent

Comme nous l'avons déjà mentionné, notre formulation actuelle du modèle présume que le portefeuille que nous construisons est efficient au sens qu'il présentera un compromis moyenne-variance optimal. Le hic, c'est qu'aucun terme dans notre critère d'apprentissage n'assure de *proximité* entre le nouveau portefeuille suggéré et l'ancien. Toutefois, soumise à de faible frais, cette négligeance ne semble pas affecter les performances du modèle outre mesure.

En finance, une mesure très utilisée pour quantifier ladite proximité est le **taux de roulement**[‡]. Dans le contexte d'une prise de décision mensuelle, par exemple, on calculera le ratio de la valeur transigée et de la valeur précédente

[‡]En anglais : *turnover*.

du portefeuille,

$$\frac{\sum_{k=0}^{N-1} |w_{kt} - w_{k(t-1)}| p_{kt}}{\sum_{k=0}^{N-1} |w_{k(t-1)}| p_{k(t-1)}}, \tag{11.6}$$

pour évaluer la proportion du portefeuille que l'on a en quelque sorte renouvellée.

Ce taux de roulement à tout intérêt à être le plus faible possible. D'abord, au niveau législatif, les gouvernements imposent des impôts plus élévés sur la portion transigée du portefeuille que sur la portion latente. Mais aussi, le taux de roulement est un bon indicateur de l'importance des frais de transactions encourus par une stratégie.

La figure (11.9) présente le taux de roulement de notre prédicteur sur l'ensemble de test. Le taux moyen est de 0.47, ce qui signifie essentiellement que, chaque mois, notre portefeuille est presqu'à moitié renouvellé. Quoi qu'il en soit, avec frais de transaction sous la barre des 0.1%, ce roulement ne peut occasioner que de très faibles pertes.

Si les transaction étaient plus onéreuses, les pertes pourraient aller jusqu'à anéantir les profits que notre modèle réaliserait par ailleurs. Alors, pour diminuer le taux de roulement de notre portefeuille, nous pourrions entres autres appliquer des règles sur le rebalancement. Par exemple, nous pourrions modifier l'ancienne position pour la nouvelle position suggérée que si la différence entre les deux était plus grande qu'un certain $\delta \in \mathbb{N}$. Ce faisant, il demeure toutefois que nous optimiserions notre modèle sous une erreur empirique qui ne se soucierait en rien des décisions passées.

Dans ce contexte, une avenue intéressante serait donc d'incorporer à notre utilité régularisée un terme qui pressente les frais de transaction, par exemple

$$\eta \sum_{k=0}^{N-1} (w_{ks} - w_{k(s-1)})^2 p_{ks}^2 \tag{11.7}$$

où η serait un nouvel hyperparamètre symbolisant la sensibilité de l'utilité au frais de transaction. Toujours est-il que, pour que cette approche porte

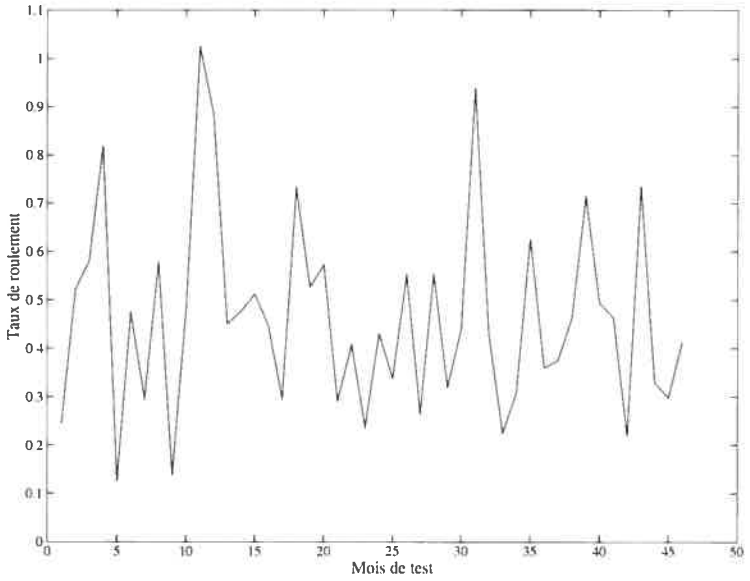

Figure 11.9 – *Taux de roulement du portefeuille sur l'ensemble de test. Le taux moyen est de 0.47, mais certain mois il dépasse même 1! Un tel phénomène est possible, même pour un portefeuille de valeur constante, étant donné le concept de vente à découvert (§ 8.1.2). En effet, si l'on possède 6 unités d'un contrat, il est tout a fait possible d'en vendre 12, ce qui occasionne un taux de roulement de 2 dans cette position, et ce sans changer la valeur investie dans ce bien,* $|w_k|p_k$.

fruit, il faudrait augmenter l'information fournie au noyau pour qu'elle soit informative quant aux décisions passées du modèle.

Chapitre 12

Conclusion

Les résultats obtenus jusqu'à maintenant nous permettent de conclure que nous avons franchi une première étape quant à l'application des méthodes à noyaux à la prise de décision sur des marchés boursiers. En effet, notre modèle a la capacité de saisir l'information sur le marché et d'en prédire un portefeuille présentant un excellent compromis moyenne-variance, même en présence de frais de transaction.

Il y a place à l'amélioration, il va s'en dire. Le modèle tel que présenté semble sensible aux changements abruptes des tendances sur le marché. Mais rappelons que nous n'avons élaboré qu'une très simple erreur empirique. Nous recherchions la convexité, pour s'assurer de trouver un minimum global, et nous nous sommes limités à l'utilisation d'une fonction objectif sans contrainte. Maintenant que nous sommes confiants du potentiel des méthodes à noyaux dans la prise de décisions financières, nous sommes à même d'envisager une erreur empirique autrement plus complexe et réaliste, qui pourrait permettre, par exemple, de mieux anticiper l'effet des frais de transaction.

Enfin, une fois ces technicalités maîtrisées dans un cadre de prises de décisions mensuelles, nous envisageons de porter le modèle vers une allocation quotidienne des ressources. Quoi qu'il en soit, le défi sera de taille. Faut-il se le rappeler, nous attaquons des données dont le la non-stationarité n'est que plus nuisible sur des périodes plus courtes. Dans cette optique, l'étude de

comités d'experts, formés de notre modèle, sous plusieurs configurations d'hyperparamètres, aussi bien que de modèles économiques classiques n'en sera que plus justifiée.

De grand défis donc, mais les accomplissements présentés dans ce mémoire, autant au niveau théorique, avec la version vectorielle du théorème du représentant, qu'au niveau pratique, avec l'obtention de performances encourageantes sous une approche minimaliste, ces accomplissements, bref, nous permettent de s'engager, sinon avec confiance, avec espoir dans l'étude desdits problèmes.

Références

AIZERMAN, M. A., . M. BRAVERMAN et L. I. ROZONOÉR (1964), « Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning », *Automation and Remote Control*.

AMIHUD, Y., J. C. BENT et H. MENDELSON (1992), « Further Evidence of Risk-Return Relationship », Rapport technique, Graduate School of Business, Stanford University.

BAKER (1977), *The Numerical Treatment of Integral Equations*, Oxford : Clarendon Press.

BELKIN, M. et P. NIYOGI (2003), « Using manifold structure for partially labeled classification », *Advances in Neural Information Processing Systems 15*, Cambridge, MA, MIT Press,

BENGIO, Y. (1997), « Using a Financial Training Criterion Rather than a Prediction Criterion », Rapport technique, Université de Montréal.

BENGIO, Y., O. DELALLEAU et N. LE ROUX (2004), « Efficient Non-Parametric Function Induction in Semi-Supervised Learning », Rapport technique, Département d'Informatique et Recherche Opérationnelle, Université de Montréal.

BENGIO, Y., O. DELALLEAU, N. LE ROUX, J.-F. PAIEMENT, P. VINCENT et M. OUIMET (2004), « Learning eigenfunctions links spectral embedding and kernel PCA », *Neural Computation submitted.*

BERG, C., J. P. R. CHRISTENSEN et P. RESSEL (1984), *Harmonic Analysis on Semigroups*, New York : Springer-Verlag.

BILLINGSLEY, P. (1995), *Probability and Measure*, New York : Wiley.

BLACK, F. (1964, July), « Capital Market Equilibrium with Restricted Borrowing », *Journal of Business 45*, p. 444–455.

BLUM, A. et T. MITCHELL (1998), « Combining Labeled and Unlabeled Data with Co-training », *COLT : Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*,

BODIE, Z., A. KANE et A. J. MARCUS (2003), *Investments, Fourth Canadian Edition.* McGraw-Hill.

BOSER, B. E., I. M. GUYON et V. N. VAPNIK (1992), « A Training Algorithm for Optimal Margin Classifiers », *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, Pittsburg, PA, ACM Press,

CHAPADOS, N. (2000), « Critères d'optimisation d'algorithmes d'apprentissage en gestion de portefeuille », Master's thesis, Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, Canada.

CHAPADOS, N. et Y. BENGIO (2003, March), « Extensions to Metric-Based Model Selection », *Journal of Machine Learning Research* (3), p. 1209–1227.

CHAPELLE, O., J. WESTON et B. SCHÖLKOPF (2003), « Cluster kernels for semi-supervised learning », *Advances in Neural Information Processing Systems 15*, Cambridge, MA, MIT Press,

CONCUS, P., G. GOLUB et D. O'LEARY (1976), « Generalized Conjugate Gradient Method for the Numerical Solution of Elliptic Partial Differential Equations. », *Sparse Matrix Computations*, p. 309–332.

COX, D. et F. O'SULLIVAN (1990), « Asymptotic Analysis of Penalized Likelihood and Related Estimators », *Annals of Statistics*, p. 1676–1695.

COZMAN, F., I. COHEN et M. CIRELO (2003), « Semi-Supervised Learning of Mixture Models », *ICML'2003*,

DIAMANTARAS, K. I. et S. Y. KUNG (1996), *Principal Component Analysis Neural Networks.* Whiley Interscience.

FAMA, E. F. (1970), « Multiperiod Consumption-Investment Decisions », *American Economic Review 60*, p. 163–174.

FAMA, E. F. et K. R. FRENCH (1992), « The Cross Section of Expected Stock Returns », *Journal of Finance 47*, p. 427–466.

GINGRAS, F., Y. BENGIO et C. NADEAU (1999), « On Out-of-Sample Statistics for Financial Time-Series », Rapport technique, Université de Montréal.

GOLUB, G. H., M. HEATH et G. WAHBA (1979), « Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter », *Technometrics 21*, p. 215–223.

GOLUB, G. H. et U. VON MATT (1997), « Generalized Cross-Validation for Large-Scale Problems », *Journal of Computational and Graphical Statistics 6*(1), p. 1–34.

GOWER, J. C. (1966), « Some Distance Properties of Latent Root and Vector Methods in Multivariate Analysis », *Biometrika 53.*

HINTON, G. E. (1986), « Learning Distributed Representations of Concepts », *Eigth Annual Conference of the Cognitive Science Society*, Erlbaum, Hillsdale,

HORE, J. E. (1987), *Trading on Canadian Futures Market* (Third ed.). The Canadian Securities Institute.

HULL, J. C. (2003), *Options, Futures and Other Derivatives* (Fifth ed.). Prentice Hall.

KHOURY, N. et P. LAROCHE (1995), *Options et Contrats à terme* (Second ed.). Les Presses de l'Université Laval.

KIMELDORF, G. S. et G. WAHBA (1971), « A Correspondance Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines », *Annals of Mathematical Statistics*, p. 495–502.

KIVINEN, J., A. SMOLA et R. WILLIAMSON (2002), « Online Learning with kernels ».

KOLMOGOROV, A. N. (1941), « Stationary Sequences in Hilbert Spaces », *Moscow University Mathematics Bulletin.*

LINTNER, J. (1965, February), « The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets », *Review of Economics and Statistics 47*, p. 13–37.

LOGUE, D. E. (1994), *Warren, Gorham, and Lamont Handbook of Financial Markets*. Howard W Sams & Co.

MARKOWITZ, H. M. (1952, March), « Portfolio Selection », *The Journal of Finance 7*(1), p. 77–91.

MARKOWITZ, H. M. (1991), *Portfolio Selection : Efficient Diversification of Investments* (Second ed.). Blackwell Publishers, originally published in 1959.

MAYERS, D. (1972), « Nonmarketable Assets and Capital Market Equilibrium under Uncertainty », *Studies in the Theory of Capital Markets*.

MERCER, J. (1909), « Functions of Positive and Negative Type and their Connection with the Theory of Integral Equations », *Philosophical Transactions of the Royal Society*, p. 415–446.

MIETZNER, A., M. OPPER et W. KINZEL (1994), « Maximal stability in unsupervised learning », *Journal of Physics A : Mathematical and General 28*, p. 2785–2797.

MOORE, D. S. et G. P. MCCABE (1999), *Introduction to the Practice of Statistics* (Third ed.). W.H. Freeman and Company.

MORGENSTERN, O. et J. VON NEUMANN (1944), *Theory of Games and Economic Behavior*. Princeton University Press.

MOSSIN, J. (1966, October), « Equilibrium in a Capital Asset Market », *Econometrica 34*, p. 768–783.

RICE, J. (1994), *Mathematical Statistics and Data Analysis* (2 ed.). Duxbury Press.

ROSS, S. A. (1976a, December), « The Arbitrage Theory of Capital Asset Pricing », *Journal of Economic Theory*.

ROSS, S. A. (1976b), « Return, Risk and Arbitrage », *Risk and Return in Finance*.

ROTH, V., T. LANGE, M. BRAUN et J. BUHMANN (2002), « A Resampling Approach to Cluster Validation ».

ROWEIS, S. T. et L. K. SAUL (2000), « Nonlinear Dimensionality Reduction by Locally Linear Embedding », *Science 290*(5500), p. 2323–2326.

SAITOH, S. (1988), *Theory of Reproducing Kernels and its Application*, England : Harlow.

SARPKAYA, S. (1989), *The Money Market in Canada* (Fourth ed.). CCH Canadian.

SCHÖLKOPF, B., R. HERBRICH, A. J. SMOLA et R. C. WILLIAMSON (2001), « A Generalized Representer Theorem », *Proceedings COLT'2001*, Springer Lecture Notes in Artificical Intelligence,

SCHÖLKOPF, B. et A. J. SMOLA (2002), *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press.

SCHÖLKOPF, B., A. J. SMOLA et K.-R. MULLER (1996), « Nonlinear Component Analysis as a Kernel Eigenvalue Problem », Rapport technique, Max-Planck-Institut für biologische Kybernetik.

SCHÖLKOPF, B., A. J. SMOLA et K.-R. MULLER (1998), « Nonlinear Component Analysis as a Kernel Eigenvalue Problem », *Advances in Kernel Methods — Support Vectors Learning*, Cambridge, MA, MIT Press,

SCHÖLKOPF, B., A. J. SMOLA et K.-R. MULLER (1999), « Kernel Principal Component Analysis », *Advances in Kernel Methods — Support Vector Learning*, Cambridge, MA, MIT Press, p. 327–352.

SCHUURMANS, D. et F. SOUTHEY (2002), « Metric-based methods for adaptive model selection and regularization », *Machine Learning 48*(1), p. 51–84.

SEEGER, M. (2001), « Learning with labeled and unlabeled data », Rapport technique, Edinburgh University.

SHARPE, W. (1964, September), « Capital Asset Prices : A Theory of Market Equilibrium », *The Journal of Finance 19*(3), p. 425–442.

SMITH, B. et B. AMOAKA-ADU (1990), *Financial Canadian Corporations with Restricted Shares.* National Center for Research and Development.

STIGUM, M. (1989), *The Money Market* (Third ed.). McGraw-Hill Trade, originally published in 1983.

STONE, M. (1974), « Cross-Validatory Choice and Assessment of Statistical Predictions », *Journal of the Royal Statistical Society. Series B 36*, p. 111–147.

SZUMMER, M. et T. JAAKKOLA (2002), « Partially labeled classification with Markov random walks », *Advances in Neural Information Processing Systems 14*, Cambridge, MA, MIT Press,

TENENBAUM, J., V. DE SILVA et J. LANGFORD (2000, December), « A Global Geometric Framework for Nonlinear Dimensionality Reduction », *Science 22*(290), p. 2319–2323.

TIKHONOV, A. et V. ARSENIN (1977), *Solutions of Ill-Posed Problems*, Washington : Winston.

TORGERSON, W. (1958), *Theory and Methods of Scaling*, New York : Wiley.

TREYNOR, J. (1961), « Towards a theory of market value of risky assets », unpublished manuscript.

VAPNIK, V. N. (1995), *The Nature of Statistical Learning Theory.* Springer-Verlag.

VAPNIK, V. N. (1998), *Statistical Learning Theory.* Wiley-Interscience.

VAPNIK, V. N. et A. J. CHERVONENKIS (1968), « On the Uniform Convergence of Relative Frequencies of Events to their Probabilities », *Doklady Akademii Nauk USSR 181*(4).

VAPNIK, V. N. et A. J. CHERVONENKIS (1971), « On the Uniform Convergence of Relative Frequencies of Events to their Probabilities », *Theory of Probability and Applications 16*, p. 264–280.

WEIGEND, A. S., Y. ABU-MOSTAFA et A.-P. REFENES (Édits.) (1997), *Decision Technologies for Financial Engineering : Prooceedings of the Fourth International Conference on Neural Networks in the Capital Markets (NNCM '96).* World Scientific Publishing.

WILLIAMS, C. et M. SEEGER (2000), « The Effect of the Input Density Distribution on Kernel-Based Classifiers », *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann,

WISKOTT, L. et T. SEJNOWSKI (2002), « Slow Feature Analysis : Unsupervised Learning of Invariances. », *Neural Computation 14*(4), p. 715–770.

ZHOU, D., O. BOUSQUET, T. NAVIN LAL, J. WESTON et B. SCHÖLKOPF (2004), « Learning with local and global consistency », *Advances in Neural Information Processing Systems 16*, Cambridge, MA, MIT Press,

ZHU, X., Z. GHAHRAMANI et J. LAFFERTY (2003), « Semi-supervised learning using gaussian fields and harmonic functions », *ICML'2003*,